Benutzererwartungen

Eine interaktive Information Retrieval Studie zur Wahrnehmung von Suchergebnissen

Katrin Werner

Information Retrieval Studie Wahrnehmung interaktive Ergebnissen Benutzererwartungen

Universitätsverlag Hildesheim

Impressum

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Universitätsverlag Hildesheim
Universitätsplatz 1
31141 Hildesheim

https://www.uni-hildesheim.de/bibliothek/publizieren/open-access-universitaetsverlag/

Erstausgabe Hildesheim 2019
Redaktion, Satz und Gestaltung: Katrin Werner

Der Beitrag ist abrufbar unter:
http://dx.doi.org/10.18442/016

Zitierempfehlung:

Werner, Katrin (2019). *Benutzererwartungen: Eine interaktive Information Retrieval Studie zur Wahrnehmung von Suchergebnissen.* Hildesheim: Universitätsverlag Hildesheim.
DOI: https://dx.doi.org/10.18442/016 (Open Access).

Benutzererwartungen: Eine interaktive Information Retrieval Studie zur Wahrnehmung von Suchergebnissen

Vom Fachbereich III (Sprach- und Informationswissenschaften)
der Universität Hildesheim

zur Erlangung des Grades
einer Doktorin der Philosophie (Dr. phil.)

angenommene Dissertation von

Katrin Werner
geboren am 06.10.1982 in Kassel

Gutachter:
Prof. Dr. Christa Womser-Hacker
Prof. Dr. Thomas Mandl
Prof. Dr. Werner Greve

Tag der mündlichen Prüfung: 11.07.2018

Bibliothekssiegel: Hil 2

I have heard it said that measuring customer satisfaction cannot be very difficult. After all, you are either satisfied with the service you receive or you are not. If you get what you want you are satisfied, if you don't you are not.

P. Szwarc

Danksagung

Das Verfassen einer Promotion ist mitunter ein langer Weg, der sich anfangs nur erahnen lässt. Diese Dissertation bildet den Abschluss meiner mehrjährigen Forschung zur Untersuchung der subjektiven Wahrnehmung von Suchergebnissen. Sie bot mir die seltene Chance, sich mit einem Forschungsgegenstand so ausführlich auseinanderzusetzen, wie es im normalen Projekt- und Forschungsalltag nur schwer möglich ist. Viele Personen haben mich auf diesem Weg begleitet und mir geholfen den Fokus nicht zu verlieren. Ihnen allen möchte ich an dieser Stelle meinen Dank aussprechen.

Dieser gilt an erster Stelle meinen Doktoreltern aus der Informationswissenschaft, Christa Womser-Hacker und Thomas Mandl. Ganz besonders möchte ich mich bei ihnen für das mir entgegengebrachte Vertrauen und die Freiheit bedanken, die es mir ermöglicht haben, dieses Thema sehr selbstständig aber nie alleingelassen und mit aller notwendigen Unterstützung bearbeiten zu können. Ich bedanke mich für viele hilfreiche Diskussionen und Anregungen, die mir immer wieder neue Denkanstöße gegeben und die Entwicklung meiner Forschung maßgeblich beeinflusst haben. In gleichem Maße möchte ich mich auch bei Werner Greve für sein fortwährendes Interesse am Fortgang dieser Arbeit und die vielen methodischen Anregungen bedanken. Sein psychologischer Blickwinkel hat maßgeblich zum Gelingen der Nutzerstudien beigetragen und ich habe unsere Treffen stets als Ermutigung und Motivation empfunden.

Darüber hinaus haben viele Kollegen durch anregende Diskussionen oder praktische Hilfestellungen zum Gelingen meines Dissertationsprojekts beigetragen. Bei allen Mitarbeitern des Instituts für Informationswissenschaft und Sprachtechnologie möchte ich mich für die offenen und anregenden Gespräche im Zuge unserer Doktorandenworkshops bedanken. Danken möchte ich außerdem meinen Mitdoktoranden, Daniela Becks, Stefanie Elbeshausen, Noushin Fadaei, Matthias Görtz, Ben Heuwing, Gabriele Irle, Julia Jürgens, Fritz Kliche, Nadine Mahrholz, Julia Maria Struß und Saskia Untiet-Kepp, die nicht nur bei vielen Fragen erste Ansprechpartner waren, sondern mich auch im Rahmen der Korupuserstellung als Juroren unterstützt und meine Arbeit Korrektur gelesen haben.

Ohne Probanden wäre interaktive Information Retrieval Forschung nicht möglich. Mein besonderer Dank gilt daher allen freiwilligen Teilnehmern der drei Nutzerstudien, deren Suchverhalten ich analysieren durfte. Besonders danken möchte ich zudem Linda Achilles, die mich bei der dritten Nutzerstudie als Hiwi unterstützt hat.

Ich möchte diese Gelegenheit auch nutzen, mich bei meinen Freundinnen Karina Glahn und Ulrike Käbberich für ihre wunderbare Gastfreundschaft während meiner Pendlerzeit zu bedanken. Ich habe die gemeinsamen Abende mit Euch immer sehr genossen.

Den wichtigsten Personen in meinem Leben möchte ich zuletzt meinen tief empfunden Dank aussprechen, auch wenn sich meine Dankbarkeit eigentlich nicht in Worte fassen lässt: Meiner gesamten Familie und vor allem meinen Eltern danke ich für die entgegengebrachte Unterstützung und Motivation. Der wohl größte Dank aber gebührt meinem Mann Albert und meinen beiden Töchtern Gesa und Swantje dafür, dass sie alle Unwägbarkeiten ertragen haben, die das Verfassen eine solchen Arbeit mit sich bringt.

Zusammenfassung

Ein genaues Verständnis des Suchprozesses und der Frage, wie Benutzer darin unterstützt werden können, bessere Suchergebnisse zu erhalten, stellt eine wichtige Aufgabe der Information-Retrieval-Forschung dar. Jedoch tragen unter anderem die dynamische Natur dieses Prozesses, die aktive Beteiligung des Nutzers am Suchverlauf sowie die Kontextabhängigkeit des Relevanzbegriffs zur Komplexität dieser Fragestellungen bei. Eine besondere Herausforderung besteht daher darin, im Kontext dieses komplexen Prozesses, die genauen Wirkungsmechanismen aller beteiligten Einflussgrößen und ihre Abhängigkeiten zu isolieren. Im Rahmen dieser Arbeit wird daher die Auswirkung zweier wesentlicher Einflussfaktoren, der Systemleistung und der Nutzererwartung, auf Benutzerleistung, Relevanzwahrnehmung und Benutzerzufriedenheit untersucht, während andere Einflussgrößen statistisch kontrolliert werden. Zu diesem Zweck werden drei aufeinander aufbauende interaktive Information-Retrieval-Nutzerstudien geplant und durchgeführt, deren Daten im Anschluss quantitativ ausgewertet werden. Bezüglich der Benutzerzufriedenheit wird dabei insbesondere untersucht, ob sich das aus der Kundenzufriedenheitsforschung bekannte C/D-Paradigma, das die Entstehung von Zufriedenheit als Soll/Ist-Vergleich zwischen wahrgenommener und erwarteter Leistung begreift, auch auf den Kontext der Informationssuche übertragen lässt.

Auf Grundlage der vorliegenden Datenbasis kann zunächst gezeigt werden, dass eine direkte Korrelation zwischen der verwendeten Systemqualität und dem Relevanzempfinden der Testteilnehmer zu bestehen scheint. Dabei ist im direkten Vergleich zweier Suchsysteme mit unterschiedlicher Systemgüte die Anwendung restriktiverer Relevanzkriterien im Kontext des besseren Systems zu beobachten. Dieses Verhalten lässt sich insbesondere anhand der in dieser Arbeit eingeführten Imprecisionmaße nachweisen, die im Wesentlichen die Tendenz der Testpersonen erfassen, mit ihrem Relevanzurteil von den dem Testkorpus zugrundeliegenden Jurorenurteilen abzuweichen. Für recallorientierte Benutzerleistungsmaße lässt sich hingegen kein signifikanter Unterschied in Abhängigkeit der Systemleistung beobachten. In Bezug auf die Benutzerzufriedenheit scheint der beschriebene systembedingte Anpassungseffekt der Relevanzwahrnehmung zu einer Reduzierung des perzipierten Systemleistungsunterschieds zu führen, wodurch auch die Benutzerzufriedenheit nur eine schwache Abhängigkeit von der Systemleistung zeigt. Für die im Rahmen der interaktiven Information-Retrieval-Forschung bislang wenig beachteten Nutzererwartungen lässt sich hingegen ein qualitativ anderes Verhalten feststellen. Hier führt eine positive Grundeinstellung bezüglich des verwendeten Suchsystems zur Anwendung weniger restriktiver Relevanzkriterien, was sich schlussendlich in einer signifikant erhöhten Nutzerzufriedenheit im Vergleich zu Testpersonen mit einer niedrigen Erwartungshaltung widerspiegelt. Darüber hinaus ergeben sich für ausgewählte Leistungsmaße und Zufriedenheitsdimensionen auch Wechselwirkungen zwischen beiden Anpassungseffekten, welche darauf hindeuten, dass der systembedingte Anpassungseffekt der Relevanzwahrnehmung vornehmlich im Kontext einer hohen Erwartungs-

haltung zum Tragen kommt, weswegen im Umkehrschluss ein Einfluss der Systemgüte auf die Benutzerzufriedenheit hauptsächlich bei Probanden mit niedriger Erwartungshaltung zu beobachten ist. In diesem Sinne können also die Vorhersagen des C/D-Paradigmas, bei denen eine hohe Zufriedenheit mit dem Übertreffen der eigenen Erwartungen assoziiert ist, nicht bestätigt werden. Vielmehr scheint die aktive Beteiligung der Nutzer am Suchprozess zur Ausbildung anderer Wirkungsmechanismen zu führen, bei denen die Entstehung von Nutzerzufriedenheit stärker an eine positive Einstellung zum verwendeten Suchsystem gekoppelt ist.

Schlüsselwörter: Information Retrieval, benutzerorientierte Evaluierung, Benutzertest, Effektivitätsmaße, Systemleistung, Relevanz, Relevanzwahrnehmung, Benutzerleistung, Benutzerzufriedenheit, Sucherfolg, Zufriedenheitsforschung, Confirmation/Disconfirmation-Paradigma, Kundenzufriedenheit, Erwartungshaltung

Abstract

Understanding the search process and how users might be aided to achieve better search results is an important task in information retrieval research. However, the dynamical nature of this process, the involvement of the user as active participant and the contextuality of the concept of relevance all contribute to the complexity of the subject, which makes it increasingly difficult to single out particular causes for improved user performance or satisfaction. Following the interactive information retrieval paradigm, this dissertation focuses on the impact of two determining factors, user expectation and system performance, on user performance and user satisfaction together with their influence on the perception of relevance. These questions are addressed in a series of three user studies, the data of which is then analyzed quantitatively. Aside from the interdependencies between the dependent and independent factors, the question is raised whether the C/D-paradigm from customer-satisfaction research, that understands user-satisfaction as the result of a comparison between expected and perceived performance, is also valid in the context of the search process.

Based on the data collected during the three user studies users' perception of relevance can be shown to be correlated with the quality of the search system used. More precisely, when comparing two search systems with differing system quality, the employment of stricter relevance criteria in the case of higher system performance can be observed. This behaviour can in particular be witnessed by a new type of user-performance measures, called imprecisions, which are introduced in order to quantify the tendency of users to rate the relevance of a document in disagreement with the jurors' judgements underlying the test corpus. In contrast, recall-oriented performance measures show no significant system quality dependence. With respect to user satisfaction the system induced change in relevance perception seems to result in a reduced perceived difference in system quality, which in turn leads to only a weak dependence of user satisfaction on the quality of the search system. Concerning the impact of user expectations that have so far received little attention in the field of interactive information retrieval research an qualitatively different behaviour can be observed: In this case, a positive attitude towards the search system leads to the adoption of less restrictive relevance criteria, which finally reflects in higher user satisfaction ratings in comparison to subjects with lower expectations. Furthermore, selected performance measures and dimensions of satisfaction show interactions between system quality and user expectations, which indicate that the system induced adaptation effect for the relevance perception is most prominent in the context of high expectations. In accordance with this observation, an influence of system quality on user satisfaction can mainly be observed for participants with low expectations. In the light of these results, the predictions of the C/D-paradigm cannot be verified, since this would require that high satisfaction ratings should occur if expectations are exceeded, whereas here for low system performance high expectations lead to higher user satisfaction ratings in comparison with low expectations. Rather it seems that

the active involvement of the user in the search process might lead to a different interplay of user expectations and system quality, such that the formation of user satisfaction is more closely related to a positive attitude towards the search system used.

Keywords: information retrieval, user-based evaluation, user study, effectiveness measures, system performance, relevance, relevance perception, user performance, user satisfaction, search success, satisfaction research, confirmation/disconfirmation-paradigm, customer satisfaction, expectancy

Inhaltsverzeichnis

Abkürzungsverzeichnis

Speziell in dieser Arbeit verwendete Abkürzungen

$A1$, $A2$, $A3$ Suchaufgaben

B01,…,B18 Benutzerleistungsmaße der Gruppe der durschnittlichen Bewertungen von Dokumentenmengen

BP Benutzerprecision

BR Benutzerrecall

E01,…,E07 Frageitems zur Ermittlung der Benutzererwartung

E_H, E_N Treatmentgruppen des Faktors Erwartung: hoch (H) vs. niedrig (N)

F01,…,F26 Frageitems zur Ermittlung der Benutzerzufriedenheit

FF1, FF2, FF3 Forschungsfragen

$I_{G,G,H}$, $I_{G,G,N}$, $I_{G,S,H}$, $I_{G,S,N}$, $I_{S,G,H}$, $I_{S,G,N}$, $I_{S,S,H}$, $I_{S,S,N}$ Treatmentgruppen für Wechselwirkung der Faktoren System 1 (G,S), System 2 (G,S) und Erwartung (H,N)

$I_{G,H}$, $I_{G,N}$, $I_{S,H}$, $I_{S,N}$ Treatmentgruppen für Wechselwirkung der Faktoren System (G,S) und Erwartung (H,N)

$I_{G,P1}$, $I_{G,P2}$, $I_{G,P3}$, $I_{S,P1}$, $I_{S,P2}$, $I_{S,P3}$ Treatmentgruppen für Wechselwirkung der Faktoren System (G,S) und Aufgabenposition (P1,P2,P3)

$I_{H,P1}$, $I_{H,P2}$, $I_{H,P3}$, $I_{N,P1}$, $I_{N,P2}$, $I_{N,P3}$ Treatmentgruppen für Wechselwirkung der Faktoren Erwartung (H,N) und Aufgabenposition (P1,P2,P3)

K01,…,K12 Kovariaten

M01,…,M48 Benutzerleistungsmaße der Gruppe der Dokumentenmengen

P_1, P_2, P_3 Aufgabenpositionen

PCP Pre-Click-Precision

REL01,…,REL10 Anzahl der im Korpus enthaltenen und bei der Suche zurückgelieferten relevanten Dokumente

S01,…,S06 Benutzerleistungsmaße der Gruppe der sonstigen Maße

S05-log,…,S06-log Benutzerleistungsmaße der Gruppe der logarithmierten sonstigen Maße

S_G, S_S Treatmentgruppen des Faktors System: gut (G) vs. schlecht (S)

SK01,…SK19 Zufriedenheitsskalen

SK01-F,…SK19-F Zufriedenheitsskalen Faktorwerte

SK01-M,…SK19-M Zufriedenheitsskalen Mittelwerte

SK-A,…,SK-K Zufriedenheitsskalen EUCS-Instrument

SP Stichprobe

SP_A Gesamtstichprobe inklusive problematischer Fallgruppen

SP_B kontrollierte Stichprobe nach Ausschluss problematischer Fallgruppen

$SP_{A,M}$, $SP_{B,M}$ über Suchaufgaben gemittelte Stichproben

SP_{IZ} Teilstichprobe: Suchzeit bei jeder Aufgabe 10 Minuten

SP_{MT}, $SP_{A,MT}$, $SP_{B,MT}$ in Bezug auf Suchaufgaben balancierte Stichproben

SP_{MV} Teilstichprobe: möglicherweise versagte Manipulation der Erwartungshaltung

SP_N Teilstichprobe: von der Auswertung ausgeschlossene Datensätze (nicht verwendet)

SP_{OT}, $SP_{A,OT}$, $SP_{B,OT}$ nicht in Bezug auf Suchaufgaben balancierte Stichproben

SP_{SB} Teilstichprobe: Verwendung problematischer Suchbegriffe

SP_{TD} Teilstichprobe: möglicherweise Test durchschaut

SP_{UV} Teilstichprobe: alle problematischen Datensätze (unter Vorbehalt)

SP_{UZ} Teilstichprobe: Suchzeit geringer als 10 Minuten

V01,…,V80 Benutzerleistungsmaße der Gruppe der Verhältnismaße

Z01,…,Z28 Benutzerleistungsmaße der Gruppe der Zeitmaße

Z01-log,…,Z28-log Benutzerleistungsmaße der Gruppe der logarithmierten Zeitmaße

Informationswissenschaftliche Fachbegriffe, Leistungsmaße und Fragebögen

ASKU Kurzskala zur Erfassung allgemeiner Selbstwirksamkeitserwartungen

AvP Average Precision

BPref Binary Preference

C/D-Paradigma Confirmation/Disconfirmation-Paradigma

CG Cumulative Gain

CLEF Conference and Labs of the Evaluation Forum (früher: Cross-Language Evaluation Forum)

DCG Discounted Cumulative Gain

EFT Embedded Figures Test

EPR Enterprise-Resource-Planning

ESR Ephemeral State of Relevance

EUCS End-User Computing Satisfaction

IIR interaktives Information Retrieval

INCOBI Inventar zur Computerbildung

IR Information Retrieval

MAP Mean Average Precision

nDCG normalized Discounted Cumulative Gain

Prec Precision

$P@n$ Precision at n

Rec Recall

TREC Text REtrieval Conference

UX User Experience

Allgemeine statistische Begriffe

ANCOVA Kovarianzanalyse (Analysis of Covariances)

ANOVA Varianzanalyse (Analysis of Variances)

α Cronbach's Alpha

df Anzahl der Freiheitsgrade (degrees of freedom)

F Statistischer Testwert

KMO Kaiser-Meyer-Olkin-Koeffizient

p Wert zur Überprüfung des Signifikanzniveaus

PCA Hauptkomponentenanalyse (Principal Component Analysis)

POMP Percent of maximum possible score

QA Quadratische Mittelwertabweichung

SD Standardabweichung

VSS Very Simple Structure-Wert

In dieser Arbeit wird nach Möglichkeit eine geschlechtsneutrale Form genutzt, ansonsten findet das generische Maskulinum Anwendung. Gemeint sind, sofern nicht explizit angegeben,stets beide Geschlechter. Aus Gründen der besseren Lesbarkeit wird auf die Nennung beider Formen verzichtet.

1. Einleitung

Im Kontext der digitalen Revolution stellt der Aufstieg von Google als meistgenutzte Suchmaschine eine einzigartige Erfolgsgeschichte dar. Trotz des zweifellos zu Beginn dieses Jahrtausends vorhandenen objektiven Qualitätsvorsprungs ist es bemerkenswert, dass Google auch heute noch im Angesicht gleichwertiger Wettbewerber wie Bing weiterhin einen Marktanteil von etwa 90% in Europa erreicht[1]. Einen Erklärungsansatz dieses stabilen Vorsprungs liefern Studien zur Markenwahrnehmung von Suchmaschinen: Die Qualität identischer Suchergebnislisten wird von Internetnutzern als höher wahrgenommen, wenn sie mit dem Label *returned by Google* versehen sind (Jansen et al., 2007). Der Markenname Google wird also unabhängig von der objektiven Güte der Suchergebnisse als Qualitätssiegel empfunden, welches die individuelle Relevanzwahrnehmung beeinflusst und im Umkehrschluss die Wechselbereitschaft der Internetnutzer verringert. Im allgemeineren Kontext der Information-Retrieval-(IR)-Evaluierung und des Information-Searching-Behaviours zeigt dieses Verhalten, dass die individuelle Relevanzwahrnehmung von Suchmaschinennutzern sich nicht allein an objektiven Kriterien orientiert, sondern darüber hinaus auch subjektive und individuelle Komponenten aufweist. Eine umfassende Theorie der Informationssuche muss also danach streben, auch subjektive Faktoren, wie Erwartungen, Vorerfahrungen, Suchstrategien und das individuelle Informationsbedürfnis zu integrieren.

Historisch gesehen werden IR-Systeme mit dem Aufkommen der ersten Internetsuchmaschinen Mitte der 90er Jahre erstmals einer breiten Nutzerschaft zugänglich. Im Zuge dessen gewinnt eine intuitive und leicht verständliche Gestaltung dieser Systeme eine größere Bedeutung als im Fall einer überwiegenden Nutzung durch Experten. Auf Seiten der IR-Evaluierung erfordert dies eine Adaption und Weiterentwicklung der Evaluierungsmethoden hin zu einer stärkeren Berücksichtigung der Benutzersicht. Suchmaschinen sollen Nutzer bei ihrer Suche nach relevanten Inhalten unterstützen, indem sie auf Basis eingegebener Suchbegriffe relevante Dokumente bereitstellen. Anstelle des abstrakten Auffindens relevanter Dokumente tritt also die Idee eines erfolgreichen Suchprozesses, bei dem die Interaktion zwischen Mensch und System in den Mittelpunkt der Betrachtung rückt. Trotz dieser Erkenntnis hat sich bei der Evaluation von IR-Systemen zunächst ein systemzentriertes Bewertungsverfahren etabliert. Dabei werden automatisiert Suchanfragen an die zu testenden Systeme gestellt und die Ergebnisse anhand verschiedener Effektivitätsmaße verglichen. Reale Nutzer und pragmatische Faktoren spielen bei diesem Evaluierungsansatz keine Rolle. Studien der letzten Jahre zeigen jedoch, dass sich durch klassische Retrievaltests erhaltene Ergebnisse nicht ohne Weiteres auf reale Anwendungssituationen übertragen lassen, was schließlich zu einem wachsenden Interesse an benutzerorientierten Ansätzen geführt hat. Beide Herangehensweisen werden im Folgenden kurz beschrieben und bewertet.

Systemorientierte Experimente messen die Leistung eines Suchsystems daran, wie gut es rele-

[1] Quelle statcounter: `http://gs.statcounter.com/search-engine-market-share/desktop/europe/2016` (zuletzt geprüft am 25.10.2017)

vante Dokumente zu einer vorgegebenen Suchanfrage findet und zugleich irrelevante Dokumente zurückhält. Das Suchsystem selbst wird dabei als eine Art Blackbox begriffen, deren innerer Aufbau und Funktionsweise für die Evaluierung ausgeblendet werden (Womser-Hacker, 2004). Nach den ersten großen Retrievaltests, die mit der Cranfield-Kollektion durchgeführt wurden, wird dieses Vorgehen auch als Cranfield-Paradigma bezeichnet (Voorhees, 2002; Buckley u. Voorhees, 2005; Mandl, 2006; Mandl, 2008). Der Problematik fehlenden Nutzerinputs und subjektiver Einschätzungen der Retrievalleistung wird dabei durch einen komparativen Evaluierungsansatz Rechnung getragen. Im Kontext der großen Evaluierungsinitiativen wie TREC und CLEF liegt das Hauptaugenmerk also nicht auf der Leistung einer einzelnen Suchmaschine sondern vielmehr auf einem Vergleich verschiedener konkurrierender Systeme, sodass die Evaluierungsergebnisse „[...] im Vergleich ihre Gültigkeit bewahren, jedoch nicht als Einzelbewertung pro System valide sind." (Womser-Hacker, 2004) Ingwersen und Järvelin (2005, S. 113) charakterisieren in diesem Zusammenhang das Ziel des systemzentrierten Evaluierungsansatzes wie folgt: „[...] to develop algorithms to identify and rank a number of (topically) relevant documents for presentation, given a (topical) request. Research seeks to construct novel algorithms and systems, and to compare their performance with each other, finding ways of improving them." Zentral für diesen Ansatz ist eine unabhängige Bewertung der Relevanz der einzelnen Dokumente in Bezug auf die betrachtete Suchaufgabe. Erst wenn bekannt ist, welche Ergebnisse innerhalb einer Testkollektion mit einer Suchanfrage übereinstimmen, kann die erbrachte Leistung eines Systems ermittelt werden. Dabei wird angenommen, dass die Relevanz eines Dokuments bezüglich eines Informationsbedürfnisses unabhängig von anderen betrachteten Dokumenten ist. Die Bewertung der Relevanz erfolgt durch unabhängige Juroren nach vorgegebenen Kriterien, um eine möglichst objektive Bewertung zu erhalten. Zur Messung der Retrievalleistung können dann mit Hilfe der vergebenen Relevanzurteile abstrakte Effektivitätsmaße berechnet werden.

Obgleich dem systemorientierten Evaluierungsansatz eine maßgebliche Rolle bei der Fortentwicklung von IR-Systemen zukommt, steht letztlich immer die Frage nach dem Bezug zur realen Nutzungssituation im Raum. Fuhr (2011, S. 65) vergleicht deshalb die Rolle des Benutzers in klassischen Retrievaltests mit einem Orakel, das statische Relevanzurteile generiert. In der Realität ist der Prozess der Relevanzbeurteilung jedoch häufig deutlich komplexer und intransparenter als im systemorientierten Paradigma angenommen: „Relevance assessments are complex phenomena and cannot be represented as a static and precise relationship between documents and a user's question." (Park, 1993, S. 345) So kann es bspw. vorkommen, dass ein Dokument nur deshalb als irrelevant zurückgewiesen wird, weil sein Inhalt dem Benutzer bereits bekannt ist. Ebenso ist es möglich, dass ein Dokument abgelehnt wird, weil das eigene Vorwissen zur Relevanzbewertung nicht ausreicht, oder aber die Relevanz nur mit Hilfe weiterer Dokumente bewertet werden kann.

Diese Fragen nach der Übertragbarkeit klassischer Retrievaltests auf reale Anwendungskontexte bilden den Ausgangspunkt der interaktiven Information-Retrieval-(IIR)-Evaluation. Dabei tritt die Beobachtung des Nutzers im Suchprozess an die Stelle der auf die abstrakte Systemleistung beschränkten Betrachtungsweise des systemorientierten Paradigmas. Den methodischen Ansatzpunkt bilden dabei kontrollierte Laborstudien, mit denen der Einfluss einzelner Parameter des Suchkontexts auf das Nutzerverhalten ermittelt wird. Im Rahmen der Frage nach der Über-

tragbarkeit systemorientierter Ergebnisse auf die reale Suchsituation bedeutet dies bspw., dass verschiedene Gruppen von Testpersonen Rechercheaufgaben mit Systemen unterschiedlicher Retrievalqualität durchführen. Zur Messung der Benutzerleistung stehen auch hier verschiedene Bewertungsmaße zur Verfügung. Je nach zu untersuchender Fragestellung kommen neben objektiven Qualitätsparametern, wie Recall und Precision, vor allem auch subjektive Maße der individuellen Suchleistung zum Einsatz. Dabei zeigt sich insbesondere, dass Qualitätsunterschiede in Bezug auf systemorientierte Leistungsmaße zum Teil durch die Benutzer kompensiert werden können, sodass Systeme unterschiedlicher Retrievalqualität zu identischer Nutzerleistung führen (Turpin u. Hersh, 2001; Al-Maskari et al., 2006; Turpin u. Scholer, 2006).

Ein Grund für diese Diskrepanz kann in dem statischen Benutzermodell vermutet werden, das dem systemzentrierten Evaluierungsansatz zugrunde liegt und insbesondere die Dynamik und Kontextabhängigkeit des Nutzerverhaltens außer Acht lässt. Eine Grundannahme für viele der traditionellen Systemleistungsmaße besteht z.B. darin, dass ein Nutzer alle Dokumente einer Ergebnisliste der Reihe nach durchgeht und bewertet (Smucker u. Jethani, 2010a, S. 595). Dieses Verhalten ist in der Realität jedoch kaum zu beobachten und blendet darüber hinaus andere Möglichkeiten der Interaktion zwischen Suchsystem und Nutzer aus. Ingwersen und Järvelin (2005, S. 112) weisen in diesem Zusammenhang auf die begrenzte Aufnahmefähigkeit realer Anwender hin: „Information overflow became concrete in a new way - often a short query of one to five words returned hundreds of thousands of documents. However, searchers were mostly happy and only interested in the first page or two of retrieved links, thus voting for precision rather than recall as a desired attribute of search results.“ Typische Anwender wählen also nur wenige hochplatzierte Dokumente zur Ansicht aus, bevor weitere Suchanfragen gestellt werden oder die Suche vor Erreichen des Listenendes beendet wird. Smucker und Jethani (2010a, S. 602) stellen in Bezug auf die Übertragbarkeit systemorientierter Ergebnisse deshalb fest: „We found that when the user task and user interface better match the Cranfield-style evaluation metric, the metric better predicts human performance.“ Um der beschriebenen Konzentration auf die vorderen Rankingplätze Rechnung zu tragen, berücksichtigen alternative Systemleistungsmaße daher z.B. den zusätzlichen Aufwand, den das Auffinden eines relevanten Dokuments auf den hinteren Listenplätzen für den Nutzer bedeutet (Järvelin u. Kekäläinen, 2000; Järvelin u. Kekäläinen, 2002). Darüber hinaus kann die Kontextabhängigkeit der Nutzerreaktion aber auch zu komplexeren Verhaltensänderungen führen. Dies umfasst bspw. eine Anpassung der Suchstrategie in Abhängigkeit von der präsentierten Systemleistung, die sich etwa in einer höheren Anzahl von Suchanfragen oder einer Anpassung des Leseverhaltens widerspiegeln kann (Smith u. Kantor, 2008, S. 153). Ebenso stellt jedoch auch der Wechsel zu einer restriktiven bzw. moderateren Relevanzwahrnehmung eine mögliche Verhaltensänderung dar (Smucker u. Jethani, 2010a, S. 596).

Eine der wesentlichen Herausforderungen im Kontext benutzerorientierter Studien besteht darin, geeignete Methoden zu entwickeln, mit denen solche Verhaltensänderungen erkannt und beurteilt werden können. Eine Hauptschwierigkeit besteht hierbei in dem dynamischen Charakter interaktiver IR-Prozesse: „From the early Cranfield studies where relevance was assumed to be static for all users to the most recent efforts concerned with users' dynamically changing affective responses to information systems, the methods needed have become more numerous

and much more complex.“ (Palmquist u. Kim, 1998, S. 14) Eine weitere Schwierigkeit wird in den individuellen Ausgangssituationen der einzelnen Benutzer gesehen. Diese unterscheiden sich etwa in ihrem Vorwissen, ihrer Ausdauer sowie ihren Bedürfnissen und Erwartungen voneinander, was die Konzeption alltagsnaher Untersuchungssituationen, die von allen Beteiligten in ähnlicher Weise wahrgenommen werden, erheblich erschwert (Kelly, 2009, S. 4). Dies zeigt sich bspw. auch in den Studien von Saracevic und Kantor (1988b) zum Einfluss individueller Unterschiede auf die Relevanzbeurteilung: „It seems that different searchers for the same question more or less look for and retrieve a different portion of the file. They seem to see different things in a question and/or interpret them in a different way and as a result retrieve different items.“ (ebd., S. 203) An dieser Stelle zeigt sich, dass auch in benutzerorientierten Untersuchungen der Beurteilung von Relevanz eine wesentliche Rolle zukommt, wobei allerdings der Versuch, realistische IR-Prozesse zu simulieren und zu analysieren, im Vordergrund steht.

In Bezug auf die Relevanzwahrnehmung kommen Untersuchungen, die sich mit mehrstufigen Relevanzurteilen befassen, zu dem Schluss, dass sich das tatsächliche Bewertungsverhalten von Benutzern nur unzureichend durch die im Kontext systemorientierter Evaluierungen verwendeten binären Relevanzkriterien abbilden lässt, da subjektive Abweichungen erst durch eine genauere Differenzierung nachvollzogen werden können (Scholer et al., 2008; Scholer u. Turpin, 2008). Interessant ist in diesem Zusammenhang bspw., dass ein Großteil unter binären Kriterien als relevant bewerteter Dokumente bei einer Neubewertung unterschiedlicher TREC-Topics auf einer erweiterten Relevanzskala nur noch als geringfügig relevant eingestuft werden (Sormunen, 2002, S. 326 f.). Ergebnisse von Studien, die sich mit der Untersuchung individueller Relevanzkriterien befassen, legen außerdem den Schluss nahe, dass eine themenbezogene Betrachtung allein noch nicht ausreicht, um eine Relevanzentscheidung treffen zu können (Schamber, 1991; Park, 1993; Barry, 1994). Vielmehr sind immer auch persönliche und situative Faktoren zu berücksichtigen.

Im Zentrum des Forschungsinteresses dieser Arbeit steht daher speziell die Frage nach dem Einfluss des kognitiven Konstrukts der Erwartung auf die Qualitätswahrnehmung von Suchmaschinen. Besondere Aufmerksamkeit gilt dabei der Frage nach dem Zusammenhang zwischen Erwartungen und Zufriedenheit. Eine Reihe von Gründen lässt die Untersuchung der im Kontext des IR bislang kaum beachteten Erwartungseinflüsse interessant erscheinen. Zunächst stellen Erwartungen einen reizvollen Gegenstand für die interdisziplinäre Erforschung des Informationssuchverhaltens dar, denn sie spiegeln den normativen Standard wider, an dem die Qualität von Suchergebnissen gemessen wird. Dabei sind nicht nur Fragestellungen der Zufriedenheit im engeren Sinne berührt. Erwartungen sind unter anderem auch bei der Relevanzbewertung von Bedeutung. So können Cuadra und Katter (1967) bspw. in einer Benutzerstudie zeigen, dass die Nutzungsintention der jeweiligen Person auch in ihrer Beurteilung der gefundenen Dokumente zum Ausdruck kommt. Auch die Auswirkungen der einleitend erwähnten subjektiven Markenerwartungen, die bewusst oder unbewusst die Entscheidung der Nutzer für oder gegen eine bestimmte Suchmaschine beeinflussen, fallen in diesen Forschungsbereich. Dabei stellt sich insbesondere die Frage, ob und wie psychologische Aspekte wie das Image eines Markennamens im Rahmen benutzerorientierter Studien zum Informationssuchverhalten zu berücksichtigen sind. Eine Studie von Jansen et al. (2007, S. 2471) kann hier nachweisen, dass das Ansehen

einer Suchmaschine direkt mit der Beurteilung ihrer Suchleistung in Verbindung steht: „Based on average relevance ratings, there was a 25% difference between the most highly rated search engine and the lowest, even though search engine results were identical in both content and presentation.“ Darüber hinaus stehen Erwartungen in einem engen Zusammenhang mit Vertrauen. Aus Untersuchungen, die sich mit Vertrauenseffekten bei der Nutzung von Suchmaschinen befassen, geht hervor, dass sich eine gewisse Wiederholungserwartung einstellt, wenn Nutzer sehen, dass Suchergebnisse meistens in der Reihenfolge ihrer Relevanz angezeigt werden, sodass höher platzierte Treffer ungeachtet ihres Inhalts häufig bevorzugt werden (Joachims et al., 2005; Pan et al., 2007; Keane et al., 2008). Weiterhin spielen Erwartungen eine nicht zu unterschätzende Rolle hinsichtlich des Wechselverhaltens von Suchmaschinennutzern. Arbeiten, die in diesen Bereich fallen, können nachweisen, dass die Unzufriedenheit mit den erhaltenen Suchergebnissen den Hauptgrund für einen Systemwechsel darstellen (White u. Dumais, 2009; Guo et al., 2011).

Den theoretischen Ausgangspunkt zur Analyse des Nutzerverhaltens in Abhängigkeit von Erwartungshaltung und Systemleistung bildet in dieser Arbeit das in der Kundenzufriedenheitsforschung weit verbreitete Confirmation/Disconfirmation-(C/D)-Modell, das davon ausgeht, dass Kundenzufriedenheit durch die Bestätigung bzw. Enttäuschung von Erwartungen entsteht. Dabei liegt das Hauptaugenmerk auf der Identifikation konkreter Verhaltensstrategien, die sich aus einer Diskrepanz zwischen erlebter Nutzungserfahrung und initialer Erwartungshaltung ergeben. Der Fokus richtet sich somit zunächst nicht auf eine Bestimmung des relativen Beitrags der Erwartungshaltung an bspw. der Qualitätsbeurteilung, sondern auf die Klärung der Steuerbarkeit der Zufriedenheitsreaktion. Das Ziel besteht also nicht in erster Linie darin, den genauen proportionalen Anteil der Erwartungshaltung am Zufriedenheitsurteil zu quantifizieren, sondern das Wie und Warum ihres Einflusses auf die Qualitätsbeurteilung zu analysieren.

Dieser einleitende Überblick zum Stand der Forschung zeigt, dass die Rolle von Erwartungen im Kontext des Nutzerverhaltens und als Determinante der Nutzerzufriedenheit bisher nur vereinzelt bei der Betrachtung von Informationssuchprozessen Berücksichtigung findet. Dies bildet einen der wissenschaftlichen Ausgangspunkte der vorliegenden Arbeit, auf den in Abschnitt 1.2 näher eingegangen wird. Der folgende Abschnitt ordnet die Arbeit zunächst noch im Kontext experimenteller Studien zum Informationssuchverhalten ein.

1.1. Einordnung der Arbeit

Schwerpunktmäßig beschäftigt sich diese Arbeit mit der Frage, welche Auswirkungen unterschiedliche Systemqualitäten und Erwartungshaltungen auf die Wahrnehmung von Suchergebnissen haben können. Ein wichtiges Thema in diesem Zusammenhang ist das dynamische Anpassungsverhalten der Relevanzbeurteilung gefundener Dokumente. Ebenfalls behandelt wird aber auch die Frage nach dem tatsächlichen Mehrwert besserer Suchergebnisse für den Benutzer. Methodisch ist diese Arbeit interdisziplinär ausgerichtet und tangiert bzw. integriert Inhalte der Forschungsbereiche Information Retrieval, Entscheidungs- und Zufriedenheitsforschung, Mensch-Maschine-Interaktion sowie der experimentellen Psychologie. Die Arbeit geht von der benutzerorientierten IR-Forschung aus, überträgt ein Modell, das zur Erklärung von Kundenzufriedenheit entwickelt wurde und untersucht mit dessen Hilfe den Einfluss von Systemqualität und Benutzererwartungen auf Benutzerleistung und Benutzerzufriedenheit. Insbesondere geht

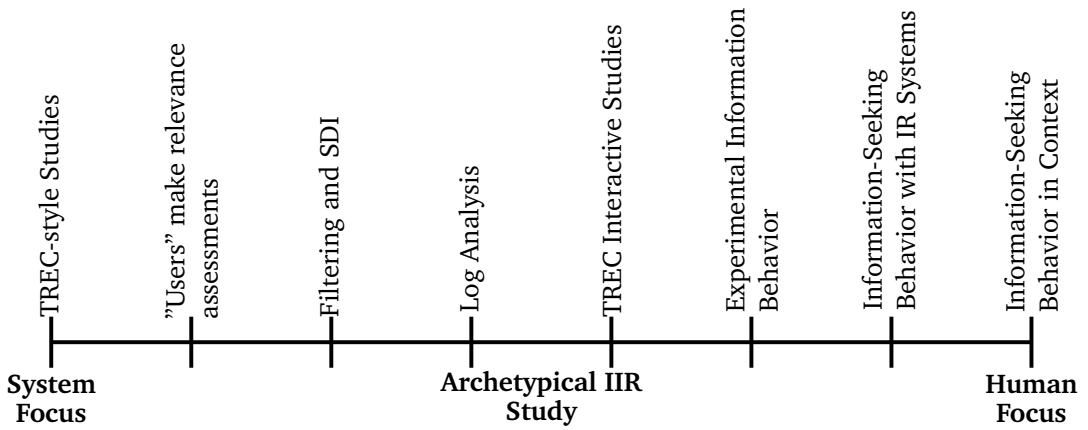


Abb. 1.1.: Typologisierung unterschiedlicher IR-Ansätze (nach Kelly, 2009, S. 10).

es dabei um die Frage, inwieweit die wahrgenommene Relevanz der Dokumente durch diese Parameter beeinflusst wird.

Im erweiterten Kontext vergleichbarer Evaluierungsstudien im IR lässt sich die vorliegende Arbeit nach Kelly (2009, S. 12 f.) als experimentelle Studie zum Informationssuchverhalten klassifizieren. Dies bedeutet, dass das Ziel einerseits darin besteht, ein möglichst realistisches Sucherlebnis zu ermöglichen, das den Testteilnehmern erlaubt, frei mit dem Suchsystem zu interagieren. Anderseits jedoch bleiben Systemqualität und Suchergebnislisten durch das Experiment kontrolliert. Folglich geht es nicht um die Frage, ob ein konkretes System besser ist als ein anderes, sondern vielmehr darum herauszufinden, wie die Probanden auf die beiden im Experiment variierten Faktoren reagieren.

Abbildung 1.1 zeigt das breite Spektrum aktueller IR-Ansätze zur Evaluation von Suchsystemen. Auf diesem Kontinuum zwischen den Extremen Systemorientierung und Benutzerorientierung positioniert sich die vorliegende Arbeit demnach auf Seite der benutzerorientierten Studien. Während in IIR-Experimenten ein spezifisches System oder eine bestimmte Systemeigenschaft mit Hilfe von Benutzern evaluiert wird, befassen sich Beobachtungsstudien zum Informationsverhalten primär mit den Informationsbedürfnissen und dem Informationsverhalten von Benutzern. Dazwischen liegen experimentelle Studien zum Informationssuchverhalten, die häufig einen bestimmten Aspekt des Suchprozesses näher betrachten. Von IIR-Studien unterscheiden sie sich vor allem aufgrund ihrer anderen Zielsetzung: „These studies are generally more interested in saying something specific about behavior, rather than on demonstrating the goodness of a particular IIR feature or system." (ebd., S. 12) Ein wesentlicher Unterschied zwischen Beobachtungsstudien und experimentellen Studien zum Informationssuchverhalten liegt in der Wahl der verwendeten Methoden. Während erstere in der Regel ein besonderes Interesse daran haben, das natürliche Verhalten der Benutzer durch den Einsatz qualitativer Methoden beobachten zu können, kommen bei letzteren oftmals die in der psychologischen Forschung üblichen Verfahren zur gezielten Kontrolle bestimmter Aspekte des Suchprozesses zur Anwendung: „For instance, the researcher might control what results are retrieved in response to a user's query or the order in which search results are presented to users." (ebd., S. 12) Die vorliegende Arbeit lässt sich somit thematisch und methodisch als experimentelle Studien zum Informationssuchverhalten einordnen.

Sowohl Kelly (ebd.) als auch Borlund (2003b) weisen darauf hin, dass die Planung und Durch-

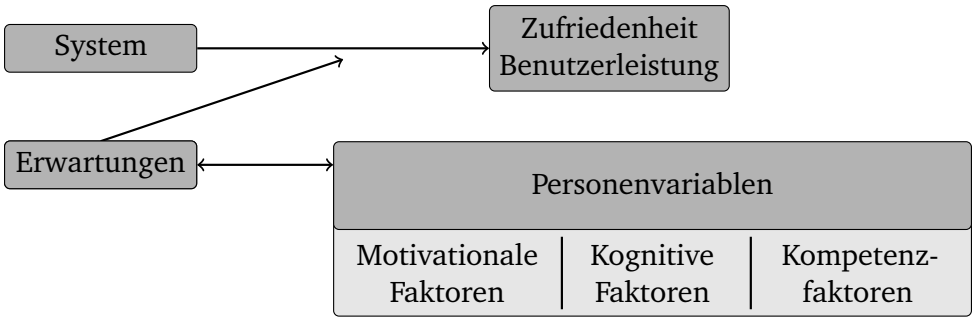


Abb. 1.2.: Theoretisches Untersuchungsmodell der Dissertation.

führung von IIR-Studien neue Herausforderungen bergen. Insbesondere verlangt die Einbeziehung realer Nutzer die Verwendung neuer Methoden aus Psychologie und Sozialwissenschaften, was zu einer natürlichen Interdisziplinarität des Forschungsgebietes führt: „The inclusion of users into any study necessarily makes IIR, in part, a behavioral science. As a result, appropriate methods for studying interactive IR systems must unite research traditions in two sciences which can be challenging.“ (Kelly, 2009, S. 4) Aus diesem Grund liegt auch in dieser Arbeit ein Schwerpunkt auf der Entwicklung methodischer Konzepte, wie dem Design der Nutzerstudien, sowie möglicher Auswertungskonzepte und -methoden für die erhobenen Nutzungsdaten.

1.2. Ziele und Forschungsfragen

Neben der Qualität eines Produktes gelten im Marketing auch Erwartungen als zentraler Einflussfaktor auf die Kundenzufriedenheit. Obwohl ein ähnlicher Zusammenhang auch für IR-Systeme vermutet wird, finden sich kaum wissenschaftliche Untersuchungen zu Erwartungen im Kontext der Suchmaschinennutzung (vgl. Abschn. 2.1.1). Diesem Zusammenhang auf den Grund zu gehen, ist daher ein zentrales Forschungsanliegen dieser Arbeit. Im Vordergrund steht dabei jedoch nicht die Bestimmung des relativen Anteils der Erwartungshaltung am Qualitätsurteil, sondern die Klärung der Steuerbarkeit der Zufriedenheitsreaktion. Es geht also weniger darum, den quantitativen Beitrag der Erwartung zu ermitteln, sondern darum, wie und warum sie einen Einfluss auf die Qualitätsbeurteilung ausübt. Die theoretische Basis der Vorhersage des Nutzerverhaltens bildet dabei das in der Kundenzufriedenheitsforschung weit verbreitete C/D-Modell (vgl. Abschn. 3.3.1.1), das die Entstehung von Kundenzufriedenheit durch die Bestätigung bzw. Enttäuschung von Erwartungen erklärt. Dieses Modell wird nun auf den Kontext der Informationssuche übertragen. Mit der Erwartung richtet diese Studie daher den Blick auf einen speziellen Teil des Suchprozesses mit dem Ziel, den Einfluss dieses Aspektes besser zu verstehen. Andere Personenvariablen, die auch Einfluss auf die Zufriedenheitsreaktion ausüben können, werden als Kovariaten berücksichtigt.

Das übergeordnete Ziel dieser Arbeit besteht darin, ein Untersuchungsmodell zu entwickeln, das die Analyse des Erwartungseinflusses unter möglichst realistischen Bedingungen gestattet, und diese dann empirisch zu überprüfen. Die theoretischen und empirischen Befunde sollen einerseits einen Beitrag zu einem verbesserten Verständnis der Qualitätswahrnehmung von Suchergebnissen leisten. Andererseits sollen Verbesserungspotenziale hinsichtlich der derzeitigen Forschungspraxis identifiziert werden.

Abbildung 1.2 zeigt das den Experimenten zugrunde gelegte Untersuchungsmodell. Ähnlich wie in anderen benutzerorientierten Studien stellt die Systemleistung eine unabhängige Variable, Benutzerzufriedenheit und -verhalten jedoch abhängige Variablen dar. Kernstück des Untersuchungsmodells und wesentliches Unterscheidungsmerkmal zu anderen Studien ist die Einbeziehung der Erwartungshaltung der Testpersonen als zusätzliche unabhängige Variable. Der Zusammenhang zwischen wahrgenommener und tatsächlicher Systemqualität wird in diesem Untersuchungsmodell nicht isoliert betrachtet. Vielmehr wird in Anlehnung an Erkenntnisse aus der Kundenzufriedenheitsforschung eine moderierende Rolle der Benutzererwartung angenommen. Zur Untersuchung dieser moderierenden Wirkung auf die Beziehung zwischen Systemqualität und Benutzerzufriedenheit werden Systemleistung und Benutzererwartung in den im Rahmen dieser Arbeit durchgeführten aufeinander aufbauenden Nutzerstudien aktiv manipuliert, während weitere Personenvariablen statistisch kontrolliert werden, um Konfundierungen zu vermeiden. Die Operationalisierung der abhängigen Variablen erfolgt über verschiedene Selbstauskunfts- und Verhaltensmaße. Dabei lassen sich die Personenvariablen in motivational-affektive, kognitive und Kompetenzfaktoren gliedern (vgl. Kap. 2). Motivationale Faktoren sind z.B. Variablen wie Interesse, intrinsische Motivation, Ausdauer und Frustrationstoleranz. Die Benutzererwartung selbst hingegen zählt zu den kognitiven Faktoren. Unter Kompetenzfaktoren fallen Variablen wie Erfahrung mit Suchmaschinen und das Vorwissen in einer Suchdomäne. Im Einzelnen werden mit Hilfe dieses Untersuchungsmodells die folgenden Fragestellungen untersucht.

Forschungsfrage 1 (FF1):
Lassen sich die Implikationen des C/D-Paradigmas auf den IR-Kontext übertragen?
Im Rahmen der ersten Forschungsfrage wird untersucht, welche Auswirkungen Benutzererwartungen auf die Qualitätswahrnehmung von Suchergebnissen haben. Leitend ist in diesem Zusammenhang die Frage nach der Übertragbarkeit der aus der Kundenzufriedenheitsforschung stammenden Diskonfirmationstheorie auf den Kontext der Informationssuche. Es wird diesbezüglich unterstellt, dass die Vorhersagen des C/D-Paradigmas auch auf den Prozess der Informationssuche mit Suchmaschinen zutreffen. Dieses Paradigma besagt, dass die Zufriedenheitsreaktion das Ergebnis eines individuellen Soll-Ist-Vergleichs ist, im Rahmen dessen die Erwartungen des Kunden der subjektiv wahrgenommenen Leistungserfüllung durch das Kaufprodukt gegenübergestellt werden. Werden die Erwartungen erfüllt oder übertroffen, ist der Kunde zufrieden. Werden die Erwartungen hingegen nicht erfüllt, d.h. die wahrgenommene Leistung ist geringer als die zuvor erwartete, so entsteht Unzufriedenheit (vgl. Abschn. 3.3.1.1).

Für eine mögliche Übertragbarkeit sprechen vor allem folgende Gründe: Vergleicht man Kauf- und Suchprozess anhand gängiger Information-Seeking-Behaviour-Modelle (Ellis, 1989; Kuhlthau, 1993a; Wilson, 1999), lassen sich strukturelle Ähnlichkeiten identifizieren. In beiden Fällen werden auf dem Weg zur Entscheidungsfindung im Wesentlichen die folgenden Phasen durchlaufen: Problem-/Bedarfserkennung, Informationssuche, Bewertung der Alternativen, Kaufentscheidung bzw. Selektion von Suchergebnissen. Der Prozess beginnt mit der Wahrnehmung eines Bedarfs. In dieser Phase wird sich der Informationssuchende seines Bedürfnisses nach einer bestimmten Information bewusst. In beiden Fällen beginnt nach der Problemerkennung die Phase der Informationssuche. In der sich anschließenden Bewertungsphase werden die Ergebnisse dieser Suche bewertet. Im Kaufprozess trifft der Kunde am Ende seine Kaufentscheidung, im

Suchprozess wählt der Nutzer eines oder mehrere Suchergebnisse aus, um diese eingehender zu betrachten. Unterscheidet man des Weiteren zwischen Kundenzufriedenheit mit Dienstleistungen und Sachleistungen, wird deutlich, dass es weitere Gemeinsamkeiten zwischen der Benutzerzufriedenheit im IR-Kontext und der Kundenzufriedenheit gibt, die die These der Übertragbarkeit bekräftigen. Gemäß Bunse (2000, S. 8) zeichnet sich Kundenzufriedenheit mit Dienstleistungen vor allem durch zwei Besonderheiten aus: „Die Einbringung eines externen Faktors [...] in den Produktionsprozess und die Immaterialität des Ergebnisses." Dabei bezieht sich der Begriff externer Faktor auf den Kunden der Dienstleistung, der anders als im Fall von Sachleistungen unmittelbar am Ergebnis der Leistung beteiligt ist und dazu beitragen kann, dass der gleiche Input des Anbieters zu individuell unterschiedlichen Ergebnissen führt. Gleiches gilt auch für die Informationssuche mit Suchmaschinen. Auch hier führt die Integration des Benutzers zu einem höheren Individualisierungsgrad. Darüber hinaus sind Produktion und Konsum nicht voneinander trennbar und die aktive Mitarbeit des Nutzers ist Voraussetzung für den Erfolg der erbrachten Leistung. Die Diskonfirmationstheorie stellt auch bei Bunse (ebd.) das Haupterklärungsmodell für die Entstehung von Kundenzufriedenheit mit Dienstleistungen dar, sodass eine Übertragung auf den Prozess der Informationssuche möglich und sinnvoll erscheint. Entsprechend der Ähnlichkeiten zum Kaufprozess sind darüber hinaus auch die Einflussfaktoren der Erwartungsbildung vergleichbar. Auch im Suchprozess spielen, wie im Kaufprozess, derzeitige Bedürfnisse und Anforderungen, bisherige Erfahrungen, die informelle Kommunikation unter Nutzern sowie die formale Anbieterkommunikation eine wichtige Rolle (vgl. Abschn. 2.1.1.1).

Natürlich erlauben diese Ähnlichkeiten allein noch nicht, diese beiden Prozesse gleichzusetzen. Dennoch ist die grundsätzliche Berücksichtigung von Erwartungen im Informationssuchprozess unverzichtbar und die Überlegung, Zufriedenheit als Ergebnis eines Erwartungs-Wahrnehmungs-Vergleichs aufzufassen, scheint ein erster lohnenswerter Ansatz zu sein.

Forschungsfrage 2 (FF2):
Führt eine höhere Systemleistung zu höherer Zufriedenheit und Benutzerleistung?
Im Kontext des benutzerorientierten Evaluierungsansatzes liegt der Fokus auf der problemorientierten Anwendung von IR-Systemen. Im Gegensatz zu klassischen Retrievaltests werden hier auch Faktoren betrachtet, die situationsbedingt sind und das Gelingen des Suchprozesses ebenfalls beeinflussen. Mit Blick auf die lange Tradition systemorientierter Forschung stellt sich jedoch auch die Frage nach der grundsätzlichen Übertragbarkeit dieser klassischen Retrievaltests auf spezifische Anwendungen und praktische Anwendungskontexte. Diesem Sachverhalt wird im Rahmen der zweiten Forschungsfrage nachgegangen, indem der Einfluss unterschiedlicher Systemgüten auf Benutzerleistung und Zufriedenheit untersucht wird. Darüber hinaus erlaubt das entwickelte Untersuchungsdesign auch eine mögliche moderierende Rolle der Erwartungshaltung auf den Einfluss der Systemgüte zu überprüfen.

Dabei liegt ein Schwerpunkt auf einer ganzheitlichen Betrachtung des Suchprozesses. Um ein möglichst natürliches Suchverhalten und somit auch ein realistisches Feedback zu fördern, werden im Zuge der Entwicklung der Experimente sowohl die Aufgaben als auch die zu ihrer Bearbeitung verwendeten Testsysteme so gestaltet, dass sie der alltäglichen Informationssuche im Internet möglichst ähnlich sind. Um gleichzeitig ein möglichst objektives Bild vom Suchprozess zu erreichen, zeichnen sich die vorliegenden Experimente außerdem durch ein hohes Maß an

Standardisierung in Durchführung und Auswertung aus. Das heißt, dass der Ablauf der Tests, die Instruktionen und die Testzeiten vorgegeben und durch einen erfahrenen Testleiter administriert werden. Auch die Auswertung erfolgt standardisiert, unabhängig vom jeweiligen Anwender und durch Mittelwertbildung über die gesamte Treatmentgruppe.

Des Weiteren richtet sich der Fokus in der dritten Nutzerstudie auf die im Rahmen der ersten beiden Experimente beobachteten Anpassung der Relevanzkriterien an die vorgefundene Systemqualität (vgl. Abschn. 5.4.2 u. 6.4.3). Um eine detailliertere Analyse dieses Effekts zu ermöglichen, wird im dritten Experiment eine feinere Relevanzskala zugrunde gelegt, mit deren Hilfe auch Aussagen über die Richtung der Anpassung möglich werden.

Forschungsfrage 3 (FF3):
Wie ändert sich die Wahrnehmung der Suchmaschinenqualität dynamisch im Suchprozess?

Nachdem in den ersten beiden Studien vor allem die grundsätzliche Wahrnehmung der Suchergebnisse im Vordergrund steht, beschäftigt sich die dritte Studie mit der Veränderung der Wahrnehmung im Prozess des Suchens. Eine Besonderheit der Suchmaschinenzufriedenheit besteht darin, dass der Nutzer als Akteur unmittelbar am Ergebnis des Suchprozesses beteiligt ist. Dies bedeutet, dass das Suchergebnis im Rahmen der Interaktion zwischen Nutzer und System zustande kommt und auf diese Weise nicht nur das Ergebnis, sondern auch sein Entstehungsprozess wahrgenommen und bewertet wird. Solche Wahrnehmungsprozesse können nur im zeitlichen Verlauf untersucht werden, weil die Verarbeitung von Informationen, deren Bewertung und schließlich die Verfestigung oder Veränderung von Relevanzkriterien erst im zeitlichen Ablauf sichtbar werden. Fragen, die in diesem Zusammenhang interessieren, betreffen z.B. die Stabilität der Wahrnehmung einer fixen Systemgüte über mehrere Suchen/Aufgaben hinweg, die Art und Weise, in welcher mit einer nicht zutreffenden Erwartungsmanipulation verfahren wird oder der Einfluss der Eigenleistung auf die Qualitätsbeurteilung.

Insbesondere die Beobachtung, dass der Einfluss der anfänglichen Erwartungsmanipulation im zweiten Experiment nach der ersten Aufgabe von der tatsächlichen Systemleistung abgelöst wird (vgl. Abschn. 6.4.5), deutet auf eine gewisse Dynamik bei der Relevanzbewertung hin. Im Rahmen der dritten Forschungsfrage soll dieser Sachverhalt näher untersucht werden. Dabei geht es wesentlich um die Frage, wie sich Benutzererwartungen im Zeitverlauf durch den Suchprozess verändern. Im Gegensatz zu den beiden anderen Forschungsfragen, bei denen jede Aufgabe für sich betrachtet wird, steht hier also die Abfolge der Suchaufgaben im Mittelpunkt.

1.3. Aufbau der Arbeit

Wie im Zusammenhang der Forschungsfragen erläutert, liegt der Fokus dieser Arbeit auf der Entwicklung eines Untersuchungsmodells zur Klärung des Erwartungseinflusses im Kontext der Suchergebniswahrnehmung, das anhand von drei Nutzerstudien evaluiert und weiterentwickelt wird. Der vorliegende Text ist dabei in neun Kapitel untergliedert. Neben einer Diskussion der theoretischen Grundlagen zur Wahrnehmung und Wahrnehmungsverarbeitung von Suchergebnissen liegt der Schwerpunkt somit auf der Darstellung der im Rahmen der Arbeit durchgeführten Experimente.

Nach der Einleitung im ersten Kapitel wird im zweiten Kapitel zunächst gezeigt, dass die Wahr-

nehmung von Suchergebnissen durch eine Vielzahl von Einflussfaktoren bestimmt ist, die in vier Unterkapiteln behandelt werden. In Abschnitt 2.1 werden zunächst die kognitiven Faktoren beschrieben, die einen Einfluss auf die Wahrnehmung der Suchmaschinenqualität haben. Kognitive Faktoren umfassen die im Gedächtnis gespeicherten Wissens- und Überzeugungsstrukturen, die Nutzer in ihrem Denken und Handeln leiten. Im Kontext der Informationssuche helfen sie dabei, die Suche zu strukturieren und gezielter zu Ergebnissen zu kommen. Ein besonderes Gewicht wird in diesem Unterkapitel auf den Stand der Forschung zu Erwartungen im Kontext der Informationssuche gelegt. Neben aktuellen Forschungsergebnissen werden hier auch die relevanten Theorien zur Erwartungsbildung erläutert. Des Weiteren wird der Einfluss von Vorerfahrungen auf die Wahrnehmung von Suchergebnissen diskutiert. Daran anschließend (Abschn. 2.2) werden motivationale Faktoren, wie bspw. die Suchmotivation beschrieben. Sie beeinflussen den Suchprozess, da sie das kognitive Engagement der Suchenden fördern (können). Auch beeinflussen sie direkt die Wahrnehmung. So kann bspw. das persönliche Interesse an einem Thema dazu führen, dass die dargebotenen Suchergebnisse positiver wahrgenommen werden. Das dritte Unterkapitel beschäftigt sich mit Faktoren, die den Status einer Person bestimmen (Abschn. 2.3). In diesem Abschnitt werden Studien beschrieben, die den Einfluss von Alter und Geschlecht auf die Informationssuche analysieren. In Abschnitt 2.4 schließlich geht es um situative Einflussfaktoren, d.h. externe Bedingungen, die von einer Suche zur nächsten variieren können. Ein besonderer Schwerpunkt wird hier auf die Aufgabenschwierigkeit im Kontext von IR-Studien gelegt.

Das dritte Kapitel befasst sich mit den drei primären Zielfaktoren experimenteller Studien zum Informationssuchverhalten. Dazu wird zunächst das zugrunde gelegte Konstrukt der situativen Relevanz erläutert (Abschn. 3.1), bevor anschließend die beiden abhängigen Variablen der vorliegenden Untersuchung genauer betrachtet werden. Während Abschnitt 3.2 die Übertragbarkeit systemorientierter Evaluationsstudien auf den Nutzerkontext sowie die Wahrnehmung des eigenen Sucherfolgs durch den Benutzer diskutiert, widmet sich Abschnitt 3.3 der subjektiv wahrgenommenen Systemqualität und der Benutzerzufriedenheit. Dabei wird deutlich, dass beide Sichtweisen wertvolle Beiträge für die Praxis der Informationssuche liefern. Die subjektivere, zufriedenheitsorientierte Sichtweise hilft einen Eindruck davon zu gewinnen, wie Nutzer das gesamte Sucherlebnis wahrnehmen und einschätzen. Die objektivere, leistungsbezogene Sichtweise hingegen ermöglicht speziellere Aspekte, wie bspw. die Relevanzwahrnehmung zu analysieren. Ein weiteres Ziel dieses Kapitels ist es in der Auseinandersetzung mit dem derzeitigen Forschungsstand den Aufbau der eigenen Untersuchungen zu präzisieren.

Kapitel 4 bildet die Brücke zwischen Theorie und Empirie. Unter Bezugnahme auf die in den vorherigen Teilen erarbeiteten Grundlagen werden in diesem Kapitel die im Rahmen dieser Arbeit verwendeten Methoden und Verfahren vorgestellt und erläutert. Das erste Unterkapitel „Das Laborexperiment als Forschungsdesign“ diskutiert zunächst die Vor- und Nachteile experimenteller Untersuchungen im Kontext interaktiver Retrievalstudien (Abschn. 4.1) und nimmt eine methodische Einordnung des den Experimenten zugrunde gelegten Forschungsdesigns vor. Anschließend werden die wichtigsten Planungsschritte zur Durchführung experimenteller Studien zum Informationssuchverhalten im Kontext des angestrebten Forschungsvorhabens diskutiert. Der darauf folgende Abschnitt 4.2 beschäftigt sich mit der konkreten Operationalisierung und Messung der in dieser Arbeit untersuchten Variablen. Neben einer auf die Experimente gerichteten Be-

trachtung der relevanten Ziel-, Einfluss- und Störfaktoren, werden insbesondere die gewählten Strategien zur Manipulation von Erwartungshaltung und Systemgüte diskutiert. Abschließend erläutert Abschnitt 4.3 die verwendeten statistischen Methoden.

Die Überprüfung der Forschungsfragen und Hypothesen der vorliegenden Arbeit erfolgt anhand drei aufeinander aufbauender experimenteller Studien. Die Kapitel 5 bis 7 beschreiben sowohl das methodische Vorgehen als auch die Ergebnisse dieser drei Experimente. Dazu sind alle Kapitel ähnlich aufgebaut. Zu Beginn werden die Untersuchungsziele definiert, Hypothesen aufgestellt sowie das verwendete Untersuchungsdesign erläutert. Daran anschließend werden die Ergebnisse der Untersuchung dargestellt und die Hypothesen geprüft. Das erste Experiment wird in Kapitel 5 beschrieben. Dieses Experiment steht ganz im Licht der Übertragbarkeit bestehender Forschungsergebnisse auf die benutzerbezogene IR-Forschung. Einerseits wird untersucht, inwiefern systemorientierte Evaluierungsergebnisse auch unter Einbeziehung realer Benutzer Bestand haben. Andererseits gilt es, die Gültigkeit des als Grundlage für diese Studie dienenden C/D-Paradigmas zu überprüfen. Aufbauend auf den Ergebnissen des ersten Experiments wird in Kapitel 6 das zweite Experiment zusammengefasst, in dem der Schwerpunkt zunächst noch einmal auf der Steuerbarkeit der Qualitätswahrnehmung liegt. Da die Erwartungsmanipulation im ersten Experiment keinen statistisch signifikanten Einfluss auf die untersuchten Zielgrößen besitzt, besteht ein wesentliches Ziel des zweiten Experiments darin, diese zu verstärken. Im dritten Experiment (Kap. 7) schließlich richtet sich der Blick auf die dynamische Entwicklung der wahrgenommenen Retrievalqualität. Dabei steht die Frage, wie sich der Einfluss von Benutzererwartungen im Prozess des Suchens verändert, im Mittelpunkt.

In Kapitel 8 werden die Ergebnisse der eigenen Untersuchungen nochmals resümiert, aufeinander bezogen und mit den im theoretischen Teil herausgearbeiteten Erkenntnissen verglichen. Daraus ergibt sich ein vollständigeres Verständnis der dynamischen Wahrnehmungsverarbeitung von Suchergebnissen. Im letzten Kapitel (Kap. 9) schließlich werden die Experimente nochmals kritisch hinterfragt und ihre Einschränkungen sowie Perspektiven für die weitere Forschung aufgezeigt.

2. Individuelle und situative Einflussfaktoren auf die Wahrnehmung von Suchergebnissen

„What people say isn't necessarily what they do." Mit dieser These bringen Mulder und Yaar (2007, S. 38) zum Ausdruck, dass das Nutzerverhalten oftmals nicht berechenbar ist und durch eine Vielzahl nicht vorhersehbarer oder gar beherrschbarer Einflüsse bestimmt wird. Die folgende Anekdote soll dies verdeutlichen: „When Sony was introducing the boom box, the company gathered a group of potential consumers and held a focus group on what color the new product should be: black or yellow. After some discussion among the group of likely buyers, everyone agreed that consumers would better respond to yellow. After the session, the facilitator thanked the group, and then mentioned that, as a bonus, they were welcome to take a free boom box on the way out. There were two piles of boom boxes: yellow and black. [...] Every person took a black boom box." (ebd., S. 38)

Die hier geschilderte Begebenheit zeigt, wie komplex und teilweise auch überraschend sich die Forschung in einem benutzerorientierten Evaluierungsprozess darstellen kann. Schnell wird klar, dass es im Kontext der Evaluierung von Suchsystemen nicht ausreicht, Probanden nur nach ihrer Zufriedenheit mit dem Suchprozess zu befragen, sondern dass es mindestens genauso wichtig ist, ihr tatsächliches Verhalten zu beobachten. Das Nutzererlebnis wird durch eine ganze Reihe von Faktoren beeinflusst: Erfahrungen mit einem System, Erwartungen und Wahrnehmungen hinsichtlich seiner Qualität, ebenso wie seine Eigenschaften und Funktionen, um nur einige Beispiele zu nennen. In einem experimentellen Forschungszusammenhang müssen diese Faktoren statistisch kontrolliert werden, damit sie die Nutzerreaktion nicht zusätzlich beeinflussen. Einflussfaktoren, die vom einzelnen Benutzer abhängen und in der Regel über eine längere Zeit und mehrere Suchen hinweg konstant bleiben, werden in dieser Arbeit als *individuelle Einflüsse* bezeichnet. Sie beziehen sich auf den persönlichen Hintergrund jedes einzelnen Nutzers. Grob lassen sie sich in *kognitive*, *motivationale* und *demographische* Faktoren untergliedern. Im Folgenden wird auf jeden einzelnen dieser Aspekte genauer eingegangen und ihre Bedeutung im Rahmen von IR-Evaluierungen anhand ausgewählter Studien erläutert. Darüber hinaus müssen Einflussfaktoren berücksichtigt werden, die sich aus der jeweiligen Suchsituation und der daraus resultierenden Dynamik ergeben. Diese werden als *situative Einflüsse* bezeichnet und sind als externe, situative Bedingungen definiert, die von einer Recherche zur nächsten variieren können. Eine zentrale Rolle nehmen im Kontext experimenteller Forschungsdesigns zum IIR die gewählten Testaufgaben ein. Während dieses Kapitel den Fokus auf Studien legt, die sich mit den Auswirkungen der Aufgabenschwierigkeit auf die Wahrnehmung von Suchergebnissen beschäftigen, werden situative Einflussfaktoren, die speziell mit der Untersuchungssituation zusammenhängen, in Kapitel 4 behandelt.

2.1. Kognitive Faktoren: Erwartungen und Vorerfahrungen

Kognitive Faktoren umfassen im Gedächtnis gespeicherte Wissens- und Überzeugungsstrukturen, die Menschen in ihrem Denken und Handeln leiten. Folglich spielen kognitive Faktoren auch bei der Bewertung von Suchergebnissen eine Rolle. So kann es bspw. sein, dass verschiedene Nutzer dieselbe Liste von Suchergebnissen unterschiedlich beurteilen, weil sie ihre eigene Sichtweise und Erfahrung einbringen.

2.1.1. Die Rolle von Erwartungen in der Suchergebniswahrnehmung

Die Entstehung von Erwartungen und ihr Einfluss auf Entscheidungs- und Bewertungssituationen im Rahmen der Informationssuche ist ein noch relativ unerforschtes Gebiet. Eine ausführlichere Auseinandersetzung mit dem Erwartungsbegriff findet sich hingegen in der Marketingforschung und in der Psychologie. Hier werden Erwartungen als Referenzgröße verstanden, mit der die erlebte Qualität eines Produkts oder einer Dienstleistung verglichen wird. Auch der Blick auf Studien aus der Informationssystemforschung deutet darauf hin, dass Erwartungen eine zentrale Rolle für die Bewertung der Systemgüte spielen. So sind Erwartungen in einer Befragung von Conrath und Mignen (1990) einer der am häufigsten genannten Einflussfaktoren auf die Benutzerzufriedenheit und auch Rushinek und Rushinek (1986, S. 597) finden heraus, dass Erwartungen einen starken Effekt auf die generelle Zufriedenheit mit dem System haben.

Allerdings ist die Bedeutung von Erwartungen für die Bewertung von Suchergebnissen bislang größtenteils unklar. Um diesen an sich naheliegenden Zusammenhang näher zu beleuchten, soll zunächst kurz auf die Definition von Erwartungen eingegangen werden. Darauf aufbauend werden Studien, die den Einfluss von Erwartungen auf die Benutzerzufriedenheit im Suchprozess behandeln, dargestellt. Dies sind zum einen Untersuchungen über die Bedeutung von Markenerwartungen und zum anderen Analysen über individuelle und generelle Determinanten der Benutzerzufriedenheit. Theoretische Erklärungsansätze zum Einfluss von Erwartungen auf die Zufriedenheit und deren Vor- und Nachteile werden hingegen in Kapitel 3 im Kontext der Zielfaktoren dieser Arbeit diskutiert.

2.1.1.1. Begriff und Verständnis von Erwartungen

Pragmatisch lassen sich Erwartungen als Einstellungen in Bezug auf kommende Ereignisse beschreiben. Im Alltagsleben stellt man sich mit ihrer Hilfe auf die Bewältigung vorhersehbarer Ereignisse ein, man trifft Entscheidungen und versucht mögliche Konsequenzen des eigenen Handelns vorauszusehen. Darüber hinaus können positive Erwartungen zu einem Hochgefühl führen, das sich auch auf das zukünftige Erleben auswirken kann. Dabei orientiert sich die Qualität der Erwartung meist an früher gemachten Beobachtungen und Erfahrungen. Trotz der Kürze dieser Darstellung lassen sich bereits wesentliche Erkenntnisse der Erwartungsforschung ableiten: Die Bildung von Erwartungen ist, ähnlich wie die dieser Arbeit zugrunde liegende Definition von Relevanz (vgl. Abschn. 3.1.1), situativ und personenspezifisch, d.h. Erwartungen variieren sowohl von Person zu Person als auch mit der Zeit. Erwartungen helfen Ängste oder Unsicherheiten abzubauen und bieten nicht zuletzt die Möglichkeit, zukünftige Ereignisse gedanklich vorauszuerleben.

Im Kontext der Kundenzufriedenheitsforschung stellen Erwartungen den Vergleichsstandard dar, an dem das jeweilige Produkt oder die Dienstleistung gemessen wird. Um ein positives

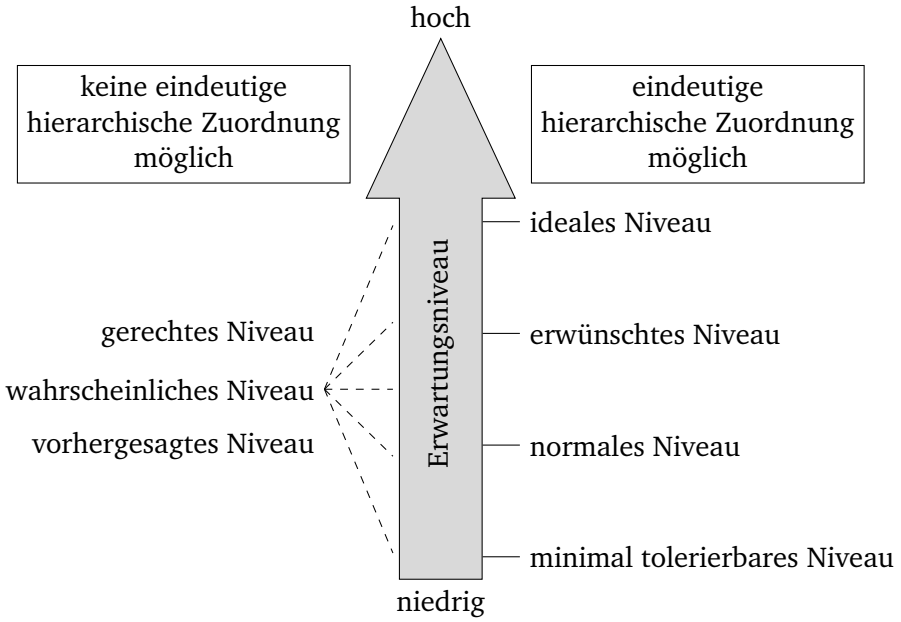


Abb. 2.1.: Hierarchie unterschiedlicher Erwartungstypen (nach Richter, 2005, S. 46).

Qualitätsurteil zu erreichen, müssen die vom Kunden gestellten Erwartungen erfüllt werden. Dabei werden zwei grundlegende Konzeptualisierungen von Erwartungen unterschieden (Bunse, 2000; Richter, 2005): Erwartungen als normativer Standard, Wunsch oder Anspruchsniveau und Erwartungen als vorhergesagter Standard, Antizipation, Prognose oder Wahrscheinlichkeitsabschätzung. Während *normative Erwartungen* den Idealzustand eines Produkts oder einer Dienstleistung beschreiben, handelt es sich bei *prädiktiven Erwartungen* um eine Form objektiver Abschätzung dessen, wie das Produkt oder die Dienstleistung aller Wahrscheinlichkeit nach sein wird. Bunse (2000, S. 20) spricht deshalb auch von Erwartungen im Sinne von „so-sollte-es-sein“ im Gegensatz zu „so-wird-es-sein“. Bunse (ebd.) merkt in diesem Zusammenhang auch an, dass prädiktive Erwartungen als Vergleichsstandard weniger geeignet sind als normative Erwartungen, da hier die persönlichen Wünsche und Bedürfnisse des Kunden außer Acht gelassen werden und folglich zwei Kunden mit unterschiedlichen subjektiven Wünschen, aber gleichen Erwartungen zu einem ähnlichen Zufriedenheitsurteil gelangen würden. Legt man stattdessen die normativen Erwartungen dieser Kunden zugrunde, so könnte sich ihr Zufriedenheitsurteil sehr wohl unterscheiden. Bunse (ebd.) argumentiert weiterhin, dass Kundenzufriedenheit ein subjektives Konstrukt darstellt und es aus diesem Grund nicht relevant ist, wie das Produkt unter objektiven Gesichtspunkten beurteilt wird. Vielmehr sei es relevant zu erfahren, wie der Kunde die Leistung persönlich empfinde oder erlebe.

Richter (2005, S. 39 ff.) nimmt neben der bereits erwähnten Einteilung in normative und prädiktive Erwartungen eine differenziertere Einteilung in sieben verschiedene Erwartungstypen vor, die in Abbildung 2.1 schematisch anhand einer hierarchischen Einordnung auf einer Leistungsskala dargestellt sind. Es fällt auf, dass bei den meisten normativen Erwartungstypen (ideales, erwünschtes, normales u. minimal tolerierbares Niveau) eine eindeutige hierarchische Einordnung möglich ist. Demgegenüber lassen sich die prädiktiven Erwartungen (wahrscheinliches u. vorhergesagtes Niveau) keinem bestimmten Leistungsniveau zuordnen, da in diesen

Fällen nicht bekannt ist, ob ein Kunde ein hohes oder aber ein niedriges Qualitätsniveau zugrundelegt. Gleiches gilt für den Erwartungstyp des gerechten Niveaus. Auch hier kann nicht objektiv nachvollzogen werden, welches Niveau zur Leistungsbeurteilung vorausgesetzt wird. Weitere Unterscheidungsdimensionen stellen das Bezugsobjekt, der Inhalt oder der Zeitaspekt des Erwartungstyps dar (Richter, 2005, S. 45). Die Unterscheidungsdimension, die das Bezugsobjekt in den Mittelpunkt stellt, fragt, ob sich die betreffende Erwartungshaltung an einen bestimmten Anbieter richtet oder ob es sich dabei um eine generelle, anbieterübergreifende Erwartungshaltung handelt. Während sich prädiktive Erwartungen in der Regel an einen bestimmten Anbieter richten, sind im normativen Fall beide Bezugsobjekte denkbar. Die nächste Unterscheidungsdimension rückt den inhaltlichen Aspekt des Erwartungstyps in den Vordergrund und fragt, ob die zugrunde gelegte Erwartungshaltung eher kognitiven oder affektiven Ursprungs ist. Auch bezogen auf diese Dimension scheinen bei normativen Erwartungen beide Aspekte eine Rolle zu spielen, während in Bezug auf prädiktive Erwartungen bisher nur kognitive Operationalisierungen bekannt sind (ebd., S. 47). Indem die letzte Unterscheidungsdimension den zeitlichen Aspekt des Erwartungstyps hervorhebt, stellt diese Dimension streng genommen nur eine Spezifizierung der beiden Grundkonzeptionalisierungen dar. Aufgrund der Tatsache, dass prädiktive Erwartungen Vorhersagen über das wahrscheinliche Leistungsniveau repräsentieren, weisen diese einen eindeutigen Zukunftsbezug auf. Im Gegensatz dazu haben normative Erwartungen eine längere zeitliche Reichweite, da sie für gewöhnlich nicht nur vor, sondern auch während und nach der Inanspruchnahme eines Produkts oder einer Dienstleistung bestand haben.

Vor dem Hintergrund, dass bei der Informationssuche auch der Nutzer selbst an der Leistungserbringung unmittelbar beteiligt ist (vgl. FF1 in Abschn. 1.2), soll an dieser Stelle noch ein weiterer Erwartungstyp erwähnt werden. Dieser geht auf die von Bandura (1986) entwickelte sozial-kognitive Lerntheorie zurück, in der davon ausgegangen wird, dass das menschliche Verhalten stark von der Einschätzung der eigenen Fähigkeiten abhängt. So belegen Studien, dass das Vertrauen in die eigenen Kompetenzen dazu beiträgt, welche Ziele eine Person für sich formuliert, wie ausdauernd sie ihre Ziele verfolgt und welche Problemlösungsstrategien im Einzelfall gewählt werden (Beierlein et al., 2013). Bandura (1986, S. 391) nennt diesen Erwartungstyp Selbstwirksamkeit: „Perceived self-efficacy is defined as people's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances. It is concerned not with the skills one has but with judgments of what one can do with whatever skills one possesses." Selbswirksamkeitserwartungen haben einen Einfluss auf zahlreiche Aspekte des alltäglichen Lebens und gelten in der Psychologie als bedeutsamer motivationaler Prädiktor für Verhalten (Beierlein et al., 2013). Der eingangs erwähnte Anteil der Eigenleistung am Suchprozess macht es somit plausibel, dass Selbstwirksamkeitserwartungen auch im Kontext der Informationssuche einen Erklärungsbeitrag leisten. Zum einen können unterschiedliche Selbstwirksamkeitserwartungen zu einem veränderten Suchverhalten führen, was in Abschnitt 2.2 erneut aufgegriffen wird, wenn es um motivationale Einflussfaktoren bei der Informationssuche geht. Zum anderen ist es jedoch auch vorstellbar, dass die Erwartung an die eigene Suchleistung den Vergleichsstandard für das eigene Zufriedenheitsurteil bedingt und somit einen unmittelbaren Einfluss auf die Zufriedenheit der Suchenden ausübt (vgl. Abschn. 3.3.1.3).

Zeithaml et al. (1993, S. 6) stellen darüber hinaus fest, dass Kunden, gleichwohl sie sich eine optimale Erfüllung ihrer Wünsche erhoffen, sehr wohl bewusst ist, dass dies nicht immer möglich ist. Demnach verfügen Kunden über einen zweiten niedrigeren Erwartungsstandard, der als *adäquates Serviceniveau* bezeichnet wird (ebd.). Die Differenz zwischen der gewünschten und dieser minimal tolerierbaren Leistung wird in der Literatur als *Indifferenz- oder Toleranzzone* bezeichnet (Woodruff et al., 1983; Zeithaml et al., 1993; Richter, 2005). Solange die erhaltene Leistung innerhalb dieser Toleranzzone liegt, betrachtet der Kunde die Leistung als zufriedenstellend. Unterschreitet sie dagegen die Mindesterwartungen, wird sich Unzufriedenheit einstellen. Ein Fokusgruppen-Teilnehmer der Studie von Zeithaml et al. (1993, S. 6) formuliert diesen Sachverhalt wie folgt: „There is a certain level of service you expect [...] as long as the service is within a certain 'window' of that level you don't complain." Für die Erwartungsmanipulation im Rahmen der in der vorliegenden Arbeit durchgeführten Experimente ist es aus diesem Grund notwendig, eine Manipulationsmethode zu finden, die den Erwartungsunterschied so einstellt, dass die Toleranzzonen der beiden experimentellen Gruppen sich möglichst nicht überschneiden (vgl. Abschn. 4.2.1.2).

Zusammenfassend lässt sich sagen, dass der Erwartungsbegriff nicht zuletzt aufgrund seiner Situations- und Personengebundenheit äußerst heterogen ist und sich deshalb im Laufe der Zeit eine Vielzahl unterschiedlicher Erwartungsbegriffe herausgebildet haben, die je nach Untersuchungsgegenstand und Forschungstradition unterschiedliche Aussagen ermöglichen. Mit Blick auf die vorliegende Arbeit bilden die unterschiedlichen Erwartungstypen eine wertvolle Hilfestellung, um die in den empirischen Arbeiten zur Informationssuche zugrunde gelegten Erwartungsbegriffe näher zu konkretisieren (vgl. Abschn. 2.1.1.3).

2.1.1.2. Der Prozess der Erwartungsbildung

Die Charakterisierung der verschiedenen Erwartungstypen hat gezeigt, dass eine Vielzahl möglicher Referenzgrößen zur Verfügung steht, wenn ein Zufriedenheitsurteil gefällt wird. Welcher Vergleichsstandard tatsächlich gewählt wird, hängt im Wesentlichen von folgenden Faktoren ab (Zeithaml et al., 1993; Scharnbacher u. Kiefer, 1996; Bunse, 2000):

Persönliche Bedürfnisse – Jeder Kunde hat eigene Vorstellungen und Voraussetzungen, die sich auf seine Wahrnehmung von Qualität auswirken. Diese können je nach Anlass und Situation unterschiedlich ausfallen. Die Suchmaschinennutzung betreffend ist anzunehmen, dass ein Dozent, der eine Literaturdatenbank verwendet, um neue Anregungen für ein Seminar zu erhalten, andere Erwartungen an ein akzeptables Suchergebnis stellen wird, als ein Student beim Verfassen seiner Abschlussarbeit. Während im ersten Fall vermutlich einige wenige Treffer zu aktuellen Entwicklungen in dem betreffenden Forschungsgebiet ausreichend sind, handelt es sich im zweiten Fall um die umfassende Bearbeitung einer wissenschaftlichen Fragestellung, für deren Beantwortung eine umfassende Literaturrecherche erforderlich ist. Neben dem Anspruch an eine angemessene Treffermenge, unterscheiden sich die hier genannten Fallbeispiele auch in ihrer Erwartung bezüglich der Aktualität der gefundenen Dokumente. Während im ersten Fall neuere Forschungsentwicklungen und weitere Forschungsbedarfe im Vordergrund stehen, geht es im zweiten Fall darum, einen möglichst vollständigen Überblick über den Stand der Forschung zu erhalten.

Vorerfahrungen – Vorerfahrungen stellen den wichtigsten Einflussfaktor der Erwartungsbildung dar (Sauerwein, 2000). Bezogen auf die Nutzung von Suchmaschinen ist davon auszugehen, dass erfahrene Nutzer andere Anforderungen an die Suchfunktionalität stellen, als unerfahrene Nutzer. So ist anzunehmen, dass erweiterte Suchmöglichkeiten, wie z.B. die Suche nach einer genauen Wortfolge, gemeinhin eher von erfahrenen Nutzern nachgefragt werden, während unerfahrene Nutzer davon keinen Gebrauch machen. Darüber hinaus können auch Erfahrungen in Bezug auf das zu bearbeitende Recherchethema die Erwartungsbildung der Nutzer beeinflussen. Verdeutlichen lässt sich dies anhand der im vorangegangenen Punkt genannten Fallbeispiele. Da der Dozent bereits über ein größeres Domänenwissen verfügt, ist zu vermuten, dass ihm die Auswahl relevanter Treffer wesentlich leichter fallen wird, als einem Studenten, der sich gerade erst in das neue Themengebiet einarbeitet.

Kommunikation über die Unternehmensleistung – Kommunikation über die Unternehmensleistung kann direkt durch den Anbieter oder indirekt durch Personen und Institutionen, die dem Unternehmen neutral gegenüberstehen, erfolgen. Im Kontext von IR-Systemen steht zu vermuten, dass insbesondere eine positive Mund-zu-Mund-Propaganda durch Freunde, Bekannte und Kollegen einen maßgeblichen Einfluss auf die Qualitätswahrnehmung der Nutzer ausübt. Um das oben genannte Beispiel fortzusetzen, würde sich der Student wahrscheinlich vor Beginn seiner Literaturrecherche bei seinen Betreuern oder Kommilitonen, über geeignete Literaturdatenbanken erkundigen.

Anzahl von Alternativen – Auch die Verfügbarkeit alternativer Angebote kann die Erwartungshaltung der Kunden steigen lassen. Dies heißt übertragen auf den Suchkontext, dass Angehörige einer Fachdisziplin mit einer Vielzahl fachspezifischer Literaturdatenbanken möglicherweise höhere Erwartungen an die Qualität der elektronischen Aufbereitung stellen, als Mitglieder einer diesbezüglich weniger gut ausgestatteten Disziplin.

Wahrgenommene Selbstwirksamkeit – Dieser Faktor umfasst die Wahrnehmung, wie stark das eigene Verhalten die erbrachte Serviceleistung beeinflussen kann. Im Kontext der Informationssuche könnte dies die Einschätzung der eigenen Fähigkeit betreffen, adäquate Suchanfragen für das gegebene Suchthema zu formulieren. Ist der Nutzer nicht davon überzeugt, eine relevante Suchanfrage gestellt zu haben – besitzt er also eine geringe Selbstwirksamkeitserwartung in Bezug auf seine Anfrage – wird er eher bereit sein, schlechte Ergebnislisten zu akzeptieren, weil er sich als mitverantwortlich begreift. Seine Toleranzzone für die erwartete Suchleistung fällt somit größer aus, als wenn er davon überzeugt wäre, eine präzise Suchanfrage gestellt zu haben.

Abbildung 2.2 fasst den Prozess der Erwartungsbildung im Hinblick auf die für den Suchprozess relevanten Faktoren nochmals graphisch zusammen: Mittig findet sich die zwischen gewünschtem und als ausreichend wahrgenommenem Service verortete Toleranzzone. Darüber hinaus sind die zuvor beschriebenen Einflussfaktoren und ihre Rolle im Erwartungsbildungsprozess dargestellt. Insbesondere lässt sich ablesen, welche Faktoren vordringlich die normativen Erwartungen ändern, welche sich stärker auf das minimale Toleranzniveau auswirken und welche

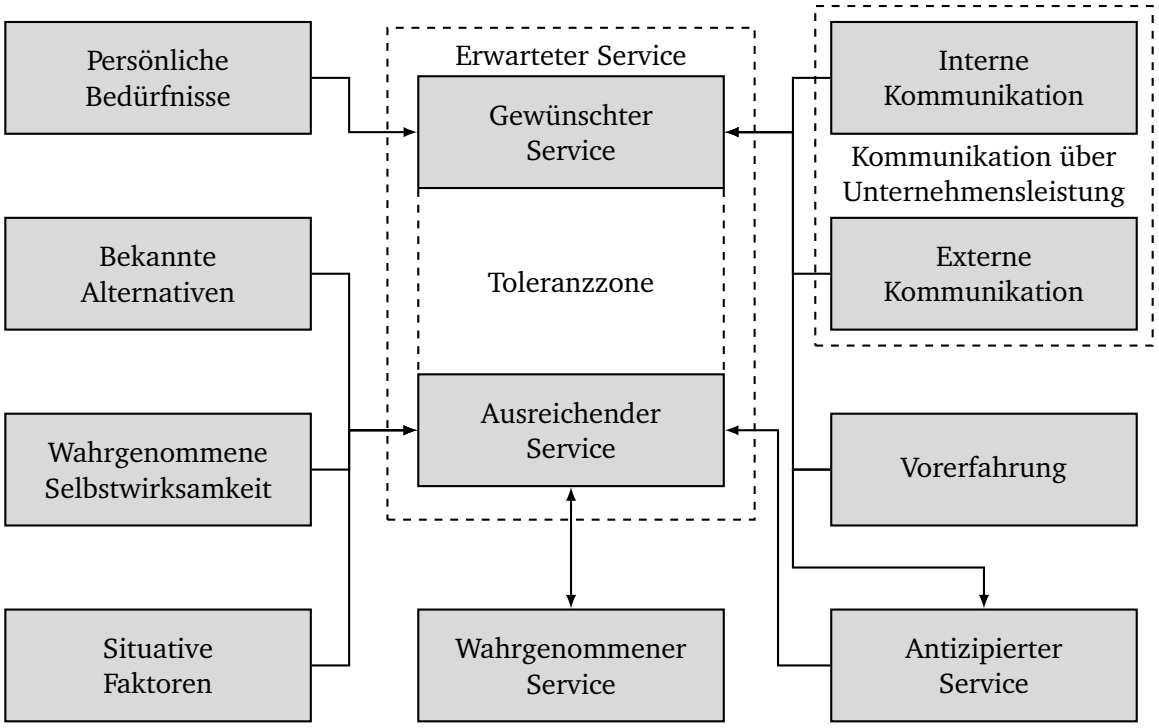


Abb. 2.2.: Determinanten der Erwartungsbildung (nach Zeithaml et al., 1993, S. 5).

Faktoren beide Ebenen beeinflussen. Während bspw. die persönlichen Bedürfnisse der Kunden zum Großteil das Anspruchsniveau verändern, wirken sich die bekannten Alternativen eher auf Erwartungen in Bezug auf einen angemessenen Service aus. Vorerfahrungen schließlich sind in der Lage, beide Enden der Toleranzzone zu beeinflussen.

Die Frage, welche Referenzgröße in einer einzelnen Situation als Vergleichsstandard herangezogen wird, unterscheidet sich sowohl interindividuell als auch intraindividuell (Richter, 2005, S. 49 ff.). Interindividuelle Unterschiede können z.B. durch verschiedene Persönlichkeitsmerkmale oder soziodemographische Faktoren entstehen. So ist anzunehmen, dass eine eher perfektionistisch veranlagte Person, die es gewohnt ist, hohe Bewertungsmaßstäbe an sich und ihre Umgebung anzulegen, auch bezogen auf die Informationssuche häufiger einen hohen Erwartungsstandard zugrundelegt. Ursachen für die intraindividuelle Variabilität bei der Wahl eines Erwartungsstandards sind z.B. in der jeweiligen Situation und der zu bewertenden Leistung zu suchen. So kann bspw. nur eine Person, die bereits Erfahrung mit einer speziellen Literaturdatenbank hat, das normale Niveau als Vergleichsstandard heranziehen. Denkbar ist ebenfalls, dass bei der Leistungsbeurteilung mehrere Standards gleichzeitig verwendet werden (ebd., S. 50 f.). Empirische Hinweise für einen Einfluss sowohl normativer als auch prädiktiver Erwartungen finden sich etwa bei Tse und Wilton (1988, S. 209). Aufgrund der Tatsache, dass diese beiden Erwartungsstandards die beiden entgegengesetzten Enden der individuellen Toleranzzone beeinflussen (vgl. Abb. 2.2), wird nochmals deutlich, dass im Rahmen dieser Arbeit darauf geachtet werden muss, die Erwartungshaltung stark zu manipulieren, um bspw. im Fall der niedrigen Erwartung den Toleranzbereich der Nutzer weit genug abzusenken.

Die vorangegangene Diskussion zeigt, dass die Erwartungsbildung vom Zusammenspiel verschiedener Faktoren abhängig ist, die ihrerseits wiederum unterschiedlich auf diesen Prozess ein-

wirken. Dabei können diese Faktoren grob danach unterschieden werden, ob sie die gewünschte Leistung, die minimal verlangte Leistung oder beide Aspekte beeinflussen. Es zeigt sich somit, dass Leistungserwartungen von Nutzern nicht punktgenau beurteilt, sondern vielmehr in Bezug auf einen Toleranzbereich evaluiert werden – eine Tatsache, die für die angestrebten Erwartungsmanipulationen im Rahmen dieser Arbeit berücksichtigt werden muss. Im Folgenden wird nun genauer auf die Rolle von Erwartungen im Kontext des Suchprozesses eingegangen.

2.1.1.3. Der Einfluss und die Entstehung von Erwartungen im Kontext der Informationssuche

Im IR-Kontext spielt die Untersuchung des Einflusses von Erwartungen auf die Zufriedenheit der Nutzer bisher nur eine untergeordnete Rolle. Im Folgenden werden daher nicht nur Studien vorgestellt, die sich direkt mit dem Einfluss von Erwartungen auf die Qualitätswahrnehmung von Suchergebnissen befassen, sondern auch Studien, die, obwohl sie sich nur am Rande mit der Erwartungsbildung von Suchmaschinennutzern beschäftigen, trotzdem zur Erschließung des Forschungsgebietes beitragen. Entsprechend der in Abschnitt 2.1.1.2 dargestellten Einflussfaktoren lassen sich die ausgewählten Forschungsarbeiten in fünf Gruppen unterteilen: Studien zu persönlichen Bedürfnissen, zu Vorerfahrungen, Studien zur Kommunikation über die Unternehmensleistung, zur Anzahl von Alternativen und zur Selbstwirksamkeit. Um der ausführlichen Darstellung in Abschnitt 2.2.2 nicht vorzugreifen, werden jedoch Studien zum Einfluss von Selbstwirksamkeitserwartungen an dieser Stelle noch nicht behandelt.

Persönliche Bedürfnisse

Persönliche Bedürfnisse und insbesondere individuelle Suchstile haben eine lange Tradition in der IR-Forschung. In diesem Kontext gibt es eine Reihe von Studien, die das individuelle Suchverhalten von Nutzern in Bezug auf unterschiedliche kognitive Stile untersuchen (Leader u. Klein, 1996; Palmquist u. Kim, 2000; White u. Drucker, 2007). Dabei bezieht sich der Terminus *kognitiver Stil* auf die Art und Weise, in der eine Person Informationen verarbeitet und organisiert. Verbreitete Anwendung fand dieser Begriff zunächst in der Lernpsychologie zur Klassifikation unterschiedlicher Lerntypen. Dabei wird der kognitive Stil einer Person als generalisiertes Merkmal oder inhärente Eigenschaft aufgefasst, die im Gegensatz zu Lerntechniken und Lernstrategien schwer zu ändern ist (Riding u. Cheema, 1991). Eines der in diesem Zusammenhang am besten evaluierten Konzepte ist die von Witkin et al. (1977) eingeführte Unterscheidung zwischen *feldabhängigen* und *feldunabhängigen* Individuen. Ursprünglich beschreiben diese Konzepte die Fähigkeit von Personen einfache geometrische Figuren in komplexen Formen wiederzuerkennen (Embedded Figures Tests - EFT). Feldunabhängige Personen schneiden hier besser ab, da sie sich auf einzelne Teile des Bildes fokussieren und das Gesamtbild ausblenden können. Folgeuntersuchungen zeigen jedoch, dass diese Disposition weitreichende Konsequenzen hat. So lässt sich bspw. die Fähigkeit feldunabhängiger Individuen, visuelle Informationen aus dem Kontext herauszulösen, auch auf ihre abstrakten Problemlösungskompetenzen übertragen. Darüber hinaus lassen sich auch soziologische Unterschiede zwischen den beiden Gruppen nachweisen. Feldabhängige Personen neigen im Lernkontext bspw. eher zu Gruppenarbeit und zeigen sogar andere Präferenzen bei der Berufswahl (Witkin, 1973).

Eine Arbeit, die den Einfluss dieser beiden kognitiven Stile auf die Suchleistung untersucht, stellt die Studie von Palmquist und Kim (2000) dar. Konkret fragen sie, welchen Einfluss der

kognitive Stil und die Sucherfahrung einer Person auf ihre Suchleistung ausübt. Sie untersuchen in ihrer Studie 48 Studierende der University of Texas in Austin. Der kognitive Stil der Untersuchungsteilnehmer wird mit Hilfe von Embedded Figures Tests ermittelt, während ihre Sucherfahrung durch einen Fragebogen erfasst wird (ebd., S. 561 f.). Als Indikatoren für die Suchleistung verwenden Palmquist und Kim (ebd., S. 562) die durchschnittliche Zeit zum Auffinden einer Information sowie die durchschnittliche Anzahl von Interaktionen mit der Universitätswebseite. Als wichtigstes Ergebnis stellte sich eine Interaktion zwischen dem kognitiven Stil und den Vorerfahrungen der Probanden heraus. Während erfahrene Nutzer keinen Unterschied hinsichtlich der Suchleistung aufweisen, benötigen unerfahrene feldabhängige Probanden mehr Zeit und Interaktionen zum Auffinden der gesuchten Information als unerfahrene feldunabhängige Teilnehmer. Spezifischer können Palmquist und Kim (ebd.) beobachten, dass unerfahrene feldabhängige Nutzer zu einer linearen Interaktion mit der Webseite neigen. Es scheint ihre implizite Erwartung zu sein, über einen zusammenhängenden Klickpfad zu der für sie relevanten Information zu gelangen. Führt solch ein Pfad in die Irre, kehren sie typischerweise direkt zur Startseite zurück anstatt im aktuellen Klickpfad nur einige Seiten zurückzugehen. In diesem Sinne verhalten sich unerfahrene feldabhängige Teilnehmer bei Palmquist und Kim (ebd.) ähnlich zu den von White und Drucker (2007) beschriebenen Navigatoren: In einer Logdatenstudie analysieren White und Drucker (ebd.) das Suchverhalten ihrer Probanden anhand ihrer Suchpfade. Zu diesem Zweck codieren White und Drucker (ebd.) die Suchpfade der Nutzer als Zeichenketten und vergleichen diese in Bezug auf ihre Levenshtein-Distanz. Es lassen sich zwei extreme Gruppen erkennen, die sich durch ihr Such- und Navigationsverhalten unterscheiden und von White und Drucker (ebd.) als *Navigatoren* bzw. *Entdecker* bezeichnet werden. Während Mitglieder der einen Gruppe bspw. konsistentere Interaktionsmuster aufweisen, weniger Rückschritte machen und eher dazu neigen, Aufgaben konsequent nacheinander zu bearbeiten, tendieren Mitglieder der anderen Gruppe dazu, mehr Suchanfragen innerhalb einer Session zu stellen, wechseln häufiger zurück und zeichnen sich insgesamt durch ein breiteres Spektrum an Interaktionsmustern aus (ebd.). Vor diesem Hintergrund könnte man sagen, dass Navigatoren die Erwartung haben, dass eine einzelne Suchstrategie für alle ihre Informationsbedürfnisse zum Erfolg führen sollte, während Entdecker ihr Suchverhalten variieren. Allerdings sollte an dieser Stelle angemerkt werden, dass White und Drucker (ebd.) keinerlei persönliche Daten ihrer Teilnehmer zur Verfügung haben. Damit bleibt bspw. offen, ob die Zugehörigkeit zu einer der beiden Gruppen mit der Vorerfahrung der Teilnehmer korreliert. Im Prinzip wäre es also möglich, dass die von White und Drucker (ebd.) beschriebene Gruppe der Navigatoren gerade mit den von Palmquist und Kim (2000) identifizierten unerfahrenen feldabhängigen Nutzern zusammenfällt.

Neben den zuvor beschriebenen inhärenten Einflüssen, wie bspw. dem kognitiven Stil des Suchenden, hat selbstverständlich auch das konkrete Informationsbedürfnis des Nutzers einen Einfluss auf seine Erwartungen. In diesem Zusammenhang ist z.B. die Arbeit von Cuadra und Katter (1967) zu nennen (vgl. Abschn. 4.2.1.2), die sich mit dem Einfluss der Instruktion von Juroren auf die Relevanzbeurteilung beschäftigt. In dieser Studie kann gezeigt werden, dass es möglich ist, die Erwartungshaltung von Testpersonen so zu manipulieren, dass diese sich je nach vorgegebener Nutzungsintention hinsichtlich ihrer Wahrnehmung und Beurteilung der Relevanz der gefundenen Dokumente unterscheiden. Vor allem in Bezug auf das in dieser Arbeit verfolgte

Forschungsziel erscheint dieses Untersuchungsergebnis interessant, da hier die Manipulation von Erwartungen eine notwendige Voraussetzung für die Durchführung der Experimente darstellt.

Abschließend lässt sich festhalten, dass persönliche Bedürfnisse und Voraussetzungen qualitativ sehr unterschiedlich auf die Erwartung Einfluss nehmen. Kognitive Stile auf der einen Seite bestimmen eher die Arbeitsweise und Art der Interaktion mit dem Suchsystem. Auf der anderen Seite stehen die Einflüsse des konkreten Suchbedürfnisses, das stärker situativ bezogen ist. Im nächsten Abschnitt wird der Einfluss des Vorwissens und der Vorerfahrung der Benutzer eingehender diskutiert.

Vorerfahrungen

Die Erfahrungen und das Vorwissen eines Nutzers können im Wesentlichen auf zwei Arten Einfluss auf den Suchverlauf nehmen: über Kompetenz- oder über Erfolgserwartungen. So wird eine Person bestimmte Suchtechniken vor allem dann ausführen, wenn sie davon ausgeht, über die entsprechenden Fähigkeiten zu verfügen (vgl. Abschn. 2.2.2). Darüber hinaus beeinflussen Erfahrungen die Bewertung der Erfolgsaussichten einer Suche. Durch die Reflexion des eigenen Suchverhaltens in früheren Nutzungskontexten entwickelt ein Benutzer bestimmte Suchstrategien und lernt, welche Strategie in welchen Situationen besonders effektiv ist (z.B. Generalisierung/Spezialisierung von Suchtermen). Angesichts einer konkreten Nutzungssituation kann dieser Benutzer dann auf unterschiedliche Verhaltensweisen zurückgreifen und diejenige auswählen, die am erfolgversprechendsten zu sein scheint. In der Literatur gibt es eine Vielzahl von Studien über den Zusammenhang von Erfahrungen und Suchverhalten. Dabei spielen im Suchkontext, wie in Abschnitt 2.1.1.2 bereits erwähnt, sowohl Sucherfahrungen als auch Erfahrungen in der Suchdomäne eine Rolle. Um die besondere Bedeutung von Erfahrungen im Suchprozess deutlich zu machen, wird dieser Einflussfaktor ausführlich in einem eigenen Unterkapitel behandelt (vgl. Abschn. 2.1.2). An dieser Stelle wird der Blick deshalb besonders auf die erwartungsbildende Funktion von Erfahrungen gerichtet. Da darüber hinaus der Einfluss von Selbstwirksamkeitserwartungen nochmals ausführlicher im Kontext motivationaler und interessensbezogener Bedingungen berücksichtigt wird (vgl. Abschn. 2.2.2), erhalten in diesem Abschnitt die Erfolgserwartungen ein größeres Gewicht.

Die erste Studie, die in diesem Rahmen vorgestellt werden soll, stammt von Cox und Fisher (2004). In einem experimentellen Untersuchungsdesign analysieren Cox und Fisher (ebd.) den Zusammenhang zwischen Erwartungen, Ergebnisqualität und Zufriedenheit. Konkret soll die Frage beantwortet werden, ob die Differenz zwischen der wahrgenommenen Qualität von Ergebnislisten und den Erfolgserwartungen in Bezug auf den Suchausgang mit der Zufriedenheit der Teilnehmer korreliert. Alle Probanden bearbeiten dieselben vier Testaufgaben in unterschiedlicher Reihenfolge, wobei die Aufgaben so konstruiert sind, dass sie in zwei Fällen hohe und in zwei Fällen niedrige Erwartungen hervorrufen sollen (ebd., S. 4). Um die Erwartungen der Probanden zu messen, werden drei Fragebogenitems eingesetzt, die folgende Aspekte der Erwartungsbildung berücksichtigen: Die erste Frage erfasst die Erwartung, überhaupt relevante Inhalte in Bezug auf die gestellte Aufgabe im Internet finden zu können. Die anderen beiden Frageitems betreffen die Einschätzungen der Testpersonen zu der Aufgabenschwierigkeit und den Suchtermen. Wenngleich dieser Punkt in der vorliegenden Studie nicht explizit untersucht wird, ist zu vermuten, dass die Vorerfahrungen der Probanden und insbesondere ihre Vorerfah-

rungen in Bezug auf die gestellte Aufgabe die Beantwortung dieser Fragen beeinflusst haben. Dies lässt sich besonders anhand des folgenden Ergebnisses illustrieren: Im Kontext der vierten Aufgabe (Find a Britsh aricraft carrier for sale), einer Aufgabe, die eigentlich zur Manipulation einer niedrigen Erfolgserwartung konstruiert ist, erwarten die meisten Probanden jedoch, dass dies möglich ist (ebd., S. 4). In der Tat berichten Cox und Fisher (ebd., S. 4), dass zum Zeitpunkt der Studie ein britischer Flugzeugträger über Ebay angeboten wird, was den meisten Studienteilnehmern aus einem aktuellen Presseartikel zu diesem Thema bekannt gewesen zu sein scheint. In einer weiteren Aufgabe (Find the 1932 school records for your grandmother, Mabel Gruenbaum) glauben nur fünf Personen, diese Information im Internet zu finden. Die meisten Testteilnehmer scheinen also ein gewisses Grundwissen darüber zu haben, wie das Internet und wie eine Suchmaschine funktioniert und welche Inhalte im Internet gefunden werden können. Die weiteren Ergebnisse dieser Untersuchung werden in Abschnitt 3.3.2.1 im Rahmen des Einflusses von Erwartungen auf die Nutzerzufriedenheit diskutiert. Eine weitere Studie, die sich mit dem Zusammenhang zwischen Erfahrungen und Erfolgserwartungen beschäftigt, ist die Studie von Kissel (1995). Ihr Ziel besteht darin, den Einfluss von Computererfahrungen auf subjektive und objektive Usability-Maße zu untersuchen. Die Vorerfahrungen der Probanden werden über einen Fragebogen erfasst. Als Testaufgaben erhalten die Teilnehmer eine Liste von Kunden-Account-Nummern, die sie über drei verschiedene Benutzeroberflächen (Kommandozeile, Menü, Listbox) wiederfinden sollen. Die Teilnehmer bewerten das Nutzererlebnis in Bezug auf Präferenz, wahrgenommene Bedienfreundlichkeit und die Erwartung, eine hohe Leistung erzielt zu haben, auf einer 7-stufigen Likert-Skala. Zusätzlich wird die Zeit zum Auffinden der Account-Nummern als objektives Usability-Maß erfasst. Die Ergebnisse der Studie zeigen, dass die Vorerfahrungen von Benutzern subjektive Produktbewertungen im Rahmen von Usability-Evaluierungen beeinflussen können (ebd., S. 285). Weiterhin lässt sich ein signifikanter Zusammenhang zwischen verschiedenen Erfahrungsvariablen und der Übereinstimmung der getesteten subjektiven und objektiven Usability-Maße nachweisen (ebd., S. 285). Auch anhand dieser Studie wird also deutlich, dass die Erfahrungen einer Person dazu beitragen, wie realistisch ihre Erwartungen ausfallen. Eine Studie von Vakkari und Hakala (2000) hingegen wählt eine zeitliche Perspektive, um den Einfluss von Erfahrungen zu untersuchen. Vakkari und Hakala (ebd., S. 546 f.) analysieren im Rahmen eines viermonatigen Seminars zur Masterarbeitsvorbereitung an der Universität Tampere das Informationsverhalten von 11 Studenten. Ziel des Seminars ist die Abgabe eines Forschungsexposés. Alle Teilnehmer verfügen über ein vergleichbares Erfahrungsniveau hinsichtlich der Suche und sind Anfänger in Bezug auf das Thema ihrer Masterarbeit. Um den Zusammenhang zwischen Änderungen bei den Relevanzkriterien und verschiedenen Phasen des Informationsprozesses untersuchen zu können, bitten Vakkari und Hakala (ebd., S. 547) ihre Studienteilnehmer einmal zu Beginn, einmal in der Mitte und einmal am Ende des Seminars eine Literatursuche durchzuführen. Sie beobachten einen Zusammenhang zwischen dem wachsenden Domänenwissen und dem Suchverhalten der Probanden. So sind die Teilnehmer mit wachsendem Themenverständis eher in der Lage, zwischen relevanten und irrelevanten Literaturquellen zu unterscheiden. Während sie zu Beginn des Informationsprozesses häufig Einführungs- und Überblicksliteratur auswählen, ist dies in späteren Prozessphasen seltener der Fall (ebd., S. 558). Ihr Anspruch und ihre Erwartung an das Suchsystem ändern sich also explizit.

Die hier diskutierten Studien stellen die Vorerfahrung der Benutzer in unterschiedlicher Weise als wichtigen Einflussfaktor auf die Erwartungsbildung heraus: Die beiden Studien von Cox und Fisher (2004) und Kissel (1995) rücken subjektive Erfolgserwartungen ins Zentrum der Betrachtung und zeigen, wie der Realitätsgrad dieser Erwartungen in Abhängigkeit von Sucherfahrungen und Vorwissen zunimmt. Die Studie von Vakkari und Hakala (2000) hebt den Einfluss von Vorerfahrungen auf individuelle Bewältigungsstrategien und Handlungsmuster hervor und zeigt, dass sich die Erwartungshaltung der Benutzer mit wachsendem Domänenwissen verändert.

Kommunikation über die Unternehmensleistung

Der Stellenwert, den immer mehr Unternehmen der eigenen Marke und deren Pflege beimessen, verdeutlicht, dass die öffentliche Wahrnehmung des Unternehmens als wichtiger Einflussfaktor auf die Kundenerwartung anerkannt ist. Auch im Kontext des IR widmen sich eine Reihe von Studien der Bedeutung der Kommunikation über die Unternehmensleistung und den damit verbundenen Vertrauenseffekten für die Nutzung von Suchmaschinen.

Jansen et al. (2007) führen in diesem Zusammenhang bspw. eine Studie zur subjektiven Markenerwartung von Suchmaschinennutzern durch. Sie stellen fest, dass ein positives Image die Kundenerwartungen steigen lässt. Die Wirkung der Markenerwartung wird über die Relevanzbewertung der Suchergebnisse erfasst, indem dieselben Ergebnisse je nach Untersuchungsbedingung in einem anderen Suchinterface (Google, MSN, Yahoo, System ohne Marke) dargestellt werden. Im Vergleich schneidet das den Teilnehmern unbekannte System bei identischen Suchergebnissen am schlechtesten ab (ebd., S. 2475). Andere Autoren untersuchen das Bewertungsverhalten von Google-Nutzern und gehen unter anderem der Frage nach, ob bei der Trefferauswahl eine Bevorzugung höher platzierter Dokumente besteht (Joachims et al., 2005; Pan et al., 2007; Keane et al., 2008). Ähnlich wie in den Studien von Eisenberg und Barry (1988) und Huang und Wang (2004) (vgl. Abschn. 4.2.3.2) wird die Reihenfolge der aus den ersten zehn Google-Treffern bestehenden Suchergebnisseiten als unabhängige Variable variiert. In der Kontrollbedingung wird die Reihenfolge der Treffer nicht verändert. In der Experimentalbedingung, die in allen drei Untersuchungen gleich ist, werden die Treffer in umgekehrter Reihenfolge angezeigt. Darüber hinaus untersuchen Joachims et al. (2007) und Pan et al. (2007) was passiert, wenn lediglich die ersten beiden Treffer vertauscht werden. Bei der Auswertung der Daten zeigt sich seitens der Probanden aller drei Studien eine Tendenz, höher platzierte Treffer auszuwählen, selbst wenn diese inhaltlich weniger relevant sind. Als Ursache kommen mehrere Möglichkeiten in Betracht. Sowohl Joachims et al. (2005) als auch Pan et al. (2007) gehen von einem Vertrauenseffekt aus, der sich ergibt, wenn Nutzer sehen, dass Suchergebnisse meistens in der Reihenfolge ihrer Relevanz angezeigt werden und ein gewisser Gewöhnungsprozess oder eine Wiederholungserwartung eintritt. Allerdings werden durch die kombinierte Analyse von Blick- und Klickdaten in der von Pan et al. (ebd.) durchgeführten Studie Anzeichen für eine zumindest implizite Wahrnehmung des Leistungsunterschieds beobachtet. Die Ergebnisse zeigen, dass Probanden, welche die umgekehrte Suchergebnisliste präsentiert bekommen, mehr Zeit für die Betrachtung der einzelnen Ergebnisseiten benötigen, eine höhere Fixationsfrequenz aufweisen und mehr Dokumente ansehen, als Probanden in der Kontrollgruppe (ebd.). Eine Studie von O'Brien und Keane (2006) hingegen stellt diese Vertrauenshypothese in Frage. O'Brien und Keane (ebd.) untersuchen den Einfluss von Reihenfolge- und Vertrauenseffekten anhand unterschiedli-

cher Trefferdarstellungen (Google-Interface vs. Linkliste). Es kann kein signifikanter Unterschied zwischen den beiden untersuchten Darstellungsweisen festgestellt werden (ebd., S. 1883). In einer Folgestudie ziehen Keane et al. (2008, S. 52) deshalb als zusätzliches Erklärungsmodell die Satisficing-Regel in Betracht (vgl. Abschn. 3.1.2). Diese besagt, dass Nutzer in bestimmten Situationen nicht nach der bestmöglichen, sondern nach einer ausreichenden Lösung mit geringem Aufwand suchen, was die beobachtete Präferenz für höher platzierter Treffer erklären würde. Um auszuschließen, dass es sich bei diesem Verhalten nicht um eine generelle Verhaltenstendenz, sondern, wie Joachims et al. (2007) und Pan et al. (2007) vermuten, um einen Vertrauenseffekt handelt, wäre es interessant, eine Studie mit einem ähnlichen Design wie in Jansen et al. (2007) durchzuführen.

Erwähnenswert ist auch eine Fragebogenstudie aus dem Bereich betrieblicher Informationssysteme, in welcher der Einfluss interner und externer Kommunikation auf die Benutzerzufriedenheit untersucht wird (Ryker et al., 1997). Basierend auf den Antworten von 252 Befragten kann gezeigt werden, dass Nutzer, deren Erwartungen zuvor durch unternehmensinterne Quellen wie Kollegen oder Servicemitarbeiter beeinflusst werden, signifikant zufriedener sind als Nutzer, deren Erwartungen durch externe Quellen wie Verkäufer oder Fernsehwerbung vorgeprägt sind. Die Autoren erklären dies mit dem höheren Realitätsgrad intern generierter Erwartungen (ebd., S. 535 f.).

In der Gesamtschau zeigen die berichteten Studien somit, dass die Markenwahrnehmung auch im Kontext der Informationssuche einen Einfluss auf die Erwartungen der Nutzer ausüben kann. Dies kann als positives Ergebnis im Sinne der in dieser Arbeit verwendeten Manipulation der Erwartungshaltung gewertet werden: Wird den Probanden eine Suchmaschine mit einem positiveren Image präsentiert, ist davon auszugehen, dass dies zu einer positiveren Erwartung der Teilnehmer führt.

Anzahl von Alternativen

Der Großteil der Studien, die das Wissen um alternative Angebote als Voraussetzung für die Bildung von Erwartungen tangieren, beschäftigen sich mit dem Loyalitäts- und Wechselverhalten von Suchmaschinenutzern. Untersuchungsgegenstand der meisten Studien ist der Systemwechsel innerhalb einer Session (Heath u. White, 2008; White u. Dumais, 2009; Guo et al., 2011). Sind Alternativen verfügbar, führen im Wesentlichen die folgenden Gründe zu einem Wechsel: Frustration oder Unzufriedenheit mit den Suchergebnissen, der Wunsch nach einer möglichst vollständigen Erschließung, die Bestätigung bereits gefundener Inhalte oder Nutzerpräferenzen in Bezug auf ein bestimmtes System (White u. Dumais, 2009, S. 89). Im Rahmen eines Feldexperiments untersuchen Guo et al. (2011) 216 Microsoft-Mitarbeiter hinsichtlich ihrer Gründe für den Wechsel zwischen verschiedenen Suchmaschinen. Ziel ist es, die Teilnehmer der Studie in einer möglichst natürlichen Nutzungssituation zu erreichen, um ein authentisches Wechselverhalten sicherzustellen. Deshalb installieren Mitarbeiter, die sich zur Teilnahme an der Studie bereit erklären, eine Browsererweiterung, die es ermöglicht einen Systemwechsel zu erkennen und in diesem Fall einen Fragebogen auf dem Bildschirm erscheinen zu lassen. Als Hauptgrund für einen Systemwechsel identifizieren Guo et al. (ebd., S. 339) die Unzufriedenheit mit der zuvor genutzten Suchmaschine. Andere Studien zum Wechselverhalten von Suchmaschinennutzern analysieren Interaktionsdaten, um Regeln zur Vorhersage eines Systemwechsels abzuleiten

(Juan u. Chang, 2005; Heath u. White, 2008; White u. Dumais, 2009; White et al., 2010; Hu et al., 2011). Juan und Chang (2005, S. 1051) führen zunächst eine Einteilung der Teilnehmer in folgende Nutzergruppen durch: Primäre System-A-Nutzer, primäre System-B-Nutzer und System-Wechsler. Die Auswertung der durchschnittlichen Sessionanzahl pro Nutzergruppe ergibt, dass Nutzer, die zwischen beiden Systemen wechseln, im Durchschnitt eine geringere Anzahl an Sessions aufweisen, was die Autoren zu folgendem Schluss veranlasst: „It suggests [that] if a search engine can make users search more, the chance of losing users to another search engine will be lower." (ebd., S. 1051) Eine ähnliche Einteilung wird von White et al. (2010) gewählt. Im Gegensatz zu Juan und Chang (2005) unterscheiden White et al. (2010) zwischen Nutzern, die nie eine andere Suchmaschine ausprobieren, System-Wechslern, die an der neuen Suchmaschine festhalten und System-Wechslern, die zwischen zwei Suchmaschinen hin- und her wechseln. Insbesondere in Bezug auf diese letzte Nutzergruppe kommt die Studie zu abweichenden Ergebnissen im Vergleich zu Juan und Chang (2005). So beobachten White et al. (2010, S. 34) bspw., dass Nutzer dieser Gruppe deutlich mehr Suchanfragen stellen. Vor dem Hintergrund einer ebenfalls von White und Drucker (2007) durchgeführten Studie zum Nutzungsverhalten erfahrener Suchmaschinennutzer (vgl. Abschn. 2.1.2), die belegt, dass zwischen Anfragefrequenz und Sucherfahrung ein deutlicher Zusammenhang besteht, kommen White et al. (2010, S. 34) zu dem Schluss, dass es sich bei dieser dritten Nutzergruppe um erfahrenere Nutzer handeln muss. White und Dumais (2009, S. 95) stellen unter anderem fest, dass einem Systemwechsel häufig mehrere Suchanfragen ohne erkennbare Trefferauswahl vorausgehen, was als Indikator für die Unzufriedenheit der Nutzer mit den präsentierten Suchergebnissen interpretiert wird. Sowohl White et al. (2010) als auch Hu et al. (2011) setzen Suchzufriedenheit mit einer Verweildauer über 30 Sekunden gleich und stellen im Zuge dessen einen Zusammenhang zwischen Nutzungsfrequenz und durchschnittlicher Zufriedenheit fest. Ein Ergebnis, das insbesondere im Hinblick auf die Untersuchungsplanung der im Rahmen dieser Arbeit durchgeführten Experimente interessant ist, betrifft die Fortdauer des korrelativen Zusammenhangs: „Essentially, we can take from this analysis that although the satisfaction ratio and rate of return fluctuate together, the effect of any fluctuations is very short-lived." (ebd., S. 1844)

Aus Sicht der Untersuchungsplanung lässt sich somit festhalten, dass in Bezug auf die an ein IR-System gerichteten Erwartungen auch eine kontextbedingte Varianz herrscht und die Zufriedenheitswahrnehmung infolgedessen nicht stabil bleibt, sondern im jeweiligen Kontext zu verorten ist. Die Messung der Zufriedenheit sollte daher möglichst zeitnah zum Stimulus erfolgen (vgl. Abschn. 4.2.2.3).

2.1.2. Die Relevanz von Erfahrungen im Kontext der Informationssuche

Nachdem im Verlauf dieses Kapitels bereits ein kurzer Einblick in die Berücksichtigung vorhandener Erfahrungen und vorhandenem Wissens als wesentliche Einflussfaktoren auf den Informationssuchprozess gegeben wurde, lohnt sich nun ein intensiverer Blick auf die Forschungsliteratur zu erfahrungsbezogenen Fragestellungen im Kontext der Informationssuche. Entsprechend der inhaltlichen Schwerpunktsetzung der Studien lässt sich dieses Unterkapitel in zwei Abschnitte gliedern: Teil eins diskutiert, in welchem Umfang sich die Beherrschung unterschiedlicher Suchstrategien und die Fähigkeit zur qualitativen Beurteilung von Information auf den Erfolg der Suche auswirken. Teil zwei hingegen untersucht, welchen Einfluss die Verfügbarkeit von

Domänenwissen auf die Relevanzbewertung besitzt. Methodische Ansätze zur Erfassung von Sucherfahrungen und Domänenwissen werden darüber hinaus in Kapitel 4 behandelt.

2.1.2.1. Der Einfluss der Suchexpertise auf das Suchverhalten

Das Wissen um die Funktionsweise von Suchmaschinen erlaubt den Nutzern, geeignete Suchstrategien zu entwickeln, um relevantes Wissen zu erschließen. Dabei umfasst erfolgreiches Recherchieren nicht nur die Verwendung sinnvoller Suchbegriffe, sondern auch die Berücksichtigung von Synonymen, Ober- und Unterbegriffen, die Nutzung von Verknüpfungsmöglichkeiten (Boolesche Operatoren) oder den Einsatz von Platzhaltern (Trunkierung/Maskierung). In diesem Abschnitt werden Studien zusammengefasst, die das Suchverhalten erfahrener und unerfahrener Benutzer vergleichen. Im Rahmen einer Logdatenstudie analysieren White und Morris (2007) das Anfrage- und Browsing-Verhalten von 188.405 Nutzern der Suchmaschinen Google, Yahoo! und MSN Search über einen Zeitraum von 13 Wochen. Als Unterscheidungskriterium zwischen erfahrenen und unerfahrenen Nutzern stellt in dieser Studie die Verwendung erweiterter Suchfunktionen dar. Dabei werden Nutzer, die im untersuchten Zeitraum mindestens eine Anfrage mit erweiterten Suchoperatoren durchführen als erfahrene Nutzer eingestuft. White und Morris (ebd., S. 258) begründen diese Entscheidung wie folgt: „The use of query operators in any queries, regardless of frequency, suggests that a user knows about the existence of the operators, and implies a greater degree of familiarity with the search system." Um das Suchverhalten dieser beiden Benutzergruppen zu vergleichen, definieren die Autoren 13 Indizes, die verschiedene Aspekte des Suchverhaltens, wie bspw. die durchschnittliche Wortanzahl pro Anfrage oder die durchschnittliche Dauer einer Suchsitzung, zusammenfassen (ebd., S. 257). Um den Erfolg der Nutzer ermitteln zu können, lassen White und Morris (ebd.) 10.680 der protokollierten Suchanfragen im Anschluss an die Datensammlung auf einer 6-stufigen Relevanzskala durch unabhängige Juroren bewerten. Die Ergebnisse der Studie zeigen eine Vielzahl signifikanter Korrelationen in die erwartete Richtung auf. Das Anfrage- und Klickverhalten der Probanden betreffend stellen White und Morris (ebd., S. 258) so z.B. fest, dass erfahrenere Nutzer weniger Suchanfragen pro Suchsession stellen, dass ihre Anfragen im Vergleich zu unerfahreneren Nutzern länger ausfallen und sie eher auch Ergebnisse anklicken, die später in der Ergebnisliste angezeigt werden. Die Analyse des Browsing-Verhaltens bestätigt die Befunde von Palmquist und Kim (2000), die zeigen, dass unerfahrene feldabhängige Probanden mehr Zeit und Interaktionen zum Auffinden der gesuchten Information benötigen (vgl. Abschn. 2.1.1.3). In beiden Untersuchungen zeichnen sich erfahrenere Nutzer durch eine zielgerichtetere Vorgehensweise aus. Dies lässt sich sowohl an der geringeren Zeit, die erfahrenere Nutzer in ihren Suchpfaden und auf einzelnen Dokumenten verbringen als auch an der niedrigeren Zahl an Wiederaufrufen bereits besuchter Seiten ablesen (White u. Morris, 2007, S. 259 f.). Diese höhere Effizienz fortgeschrittener Nutzer schlägt sich auch in einem höheren Sucherfolg nieder, der über alle in der Studie betrachteten Relevanzmaße stabil ist (ebd., S. 260 f.). Wie die Ergebnisse dieser Studie zeigen, spielen Erfahrungen eine wichtige Rolle im Suchprozess sowohl in Bezug auf die Effizienz als auch in Bezug auf die Effektivität der Suche.

Eine Reihe anderer Studien zum Einfluss von Erfahrungen mit Suchmaschinen benutzen unterschiedliche Operationalisierungen von Erfahrungen und Aufgaben sowie andere Auswertungsmethoden, kommen jedoch zu ähnlichen Ergebnissen und zeigen somit, dass die vorgestellten

Befunde als repräsentativ anzusehen sind (Gluck, 1995; Yuan, 1997; Hölscher u. Strube, 2000; Lazonder et al., 2000; Jenkins et al., 2003; Shiri u. Revie, 2003). Näher eingegangen werden soll an dieser Stelle auf die Resultate von Studien, die zusätzlich das Domänenwissen der Probanden berücksichtigen (Gluck, 1995; Hölscher u. Strube, 2000; Jenkins et al., 2003; Shiri u. Revie, 2003; Wildemuth, 2004). Besonders bedeutsam für den deutschsprachigen Raum ist in diesem Zusammenhang die Studie von Hölscher und Strube (2000), die zwei Experimente zum Einfluss von Suchexpertise durchführen. Während das erste Experiment dazu dient, ein Modell des bei Suchexperten beobachteten Anfrage- und Browsingverhaltens zu generieren, vergleichen Hölscher und Strube (ebd.) im zweiten Experiment, welches ein 2×2-Between-Subjects-Design verwendet, das Suchverhalten von 24 Probanden mit unterschiedlichem Such- und Domänenwissen. Die Webexpertise der Probanden wird in dieser Studie durch Interviews und Pretests abgefragt. Da die zu bearbeitenden Suchaufgaben *Die Europäische Währungsunion* zum Thema haben, werden in dieser Studie Studenten der Betriebswirtschaftslehre als Domänenexperten behandelt (ebd., S. 341). Generell kommt die Studie zu dem Schluss, dass eine höhere Effektivität und Effizienz nicht nur eine Frage der Sucherfahrung, sondern auch der Domänenerfahrung ist. Dabei lassen sich die zentralen Ergebnisse dieser Studie wie folgt zusammenfassen: Während Benutzer, die auf beide Erfahrungsbereiche zurückgreifen können, insgesamt gesehen am erfolgreichsten abschneiden, müssen Benutzer mit Defiziten in einem der beiden Bereiche ihre fehlende Erfahrung durch alternative Strategien kompensieren (ebd., S. 345). So fällt bei der Analyse der gestellten Suchanfragen bspw. auf, dass unerfahrene Domänenexperten ihr fehlendes Wissen über erweiterte Suchoperatoren durch eine flexiblere Suchbegriffauswahl ausgleichen: „[...] they most likely used their own terminology instead of relying on the words that were already in the original task statement. Also, more often than others they used completely different terminology from one query to the next.“ (ebd., S. 345) In Übereinstimmung mit den zuvor beschriebenen Studien stellen Hölscher und Strube (ebd., S. 343) außerdem fest, dass erfahrenere Nutzer seltener von rückwärts gerichteten Suchstrategien, wie z.B. dem Zurück-Button oder dem Wiederaufrufen bereits besuchter Seiten, Gebrauch machen (Palmquist u. Kim, 2000; White u. Morris, 2007). Erwähnenswert ist darüber hinaus, dass Hölscher und Strube (2000, S. 343) einige signifikante Interaktionen zwischen dem Domänenwissen und der Suchexpertise der Probanden nachweisen können: So ist die Wahrscheinlichkeit, tatsächlich ein Dokument von einer Ergebnisseite aufzurufen, anstatt bspw. die Suchanfrage zu reformulieren oder die Suchmaschine zu wechseln, bei Webexperten, die über ein geringes Domänenwissen verfügen, höher als bei den restlichen drei Untersuchungsgruppen. Hölscher und Strube (ebd., S. 343) vermuten, dass die Relevanzkriterien der Testpersonen in diesem Fall weniger klar formuliert sind. Diese Interpretation deckt sich mit den Befunden vergleichbarer Studien zum Einfluss des Domänenwissens (vgl. Abschn. 2.1.2.2). Eine weitere Interaktion betrifft die Gruppe der Teilnehmer, die weder über Suchexpertise noch über Domänenwissen verfügen. Es zeigt sich, dass Mitglieder dieser Gruppe ihre Suchanfragen am häufigsten umformulieren, die wenigsten Dokumente auswählen und dass darüber hinaus die meisten der ausgewählten Dokumente irrelevant sind (ebd.). Die Autoren unterziehen daraufhin die Suchanfragen dieser Benutzergruppe einer weiteren Analyse und stellen fest: „[...] that they often make only small and ineffective changes to their queries, forcing them to reiterate repeatedly.“ (ebd., S. 343) Auch Jenkins et al. (2003) verwenden

ein 2×2-Between-Subjects-Design, um den Einfluss von Suchexpertise und Domänenwissen auf das Suchverhalten von 16 Krankenschwestern zu untersuchen. Gemäß dieser medizinischen Zielgruppe ist das übergreifende Thema der Suchaufgaben Osteoporose. Eine Gruppe von acht Krankenschwestern, die ein Zentrum zu Osteoporose-Forschung besucht haben, werden in dieser Studie als Domänenexperten eingestuft (ebd., S. 69). Webexpertise definieren Jenkins et al. (ebd., S. 69) als mindestens ein Jahr Internetnutzung. Im Gegensatz zu der zuvor beschriebenen Untersuchung von Hölscher und Strube (2000) nutzen Jenkins et al. (2003) eine qualitative Analysemethode, bei der konkrete Einzelfälle betrachtet und ausgewertet werden. Im Großen und Ganzen bestätigen ihre Ergebnisse die bisher vorgestellten Befunde. Insbesondere beobachten sie die auch im Kontext der anderen Studien betonte Verhaltenstendenz eher unerfahrener Nutzer, die Suchergebnisseite als Zentrum der Suche aufzufassen, zu dem sie immer wieder zurückkehren, was Jenkins et al. (ebd.) als Breitensuche (breadth-first search) bezeichnen. Auch bestätigen sie die Tendenz eher erfahrener Nutzer, sich im Verlauf der Suche weiter von diesem Zentrum zu entfernen, was sie als Tiefensuche (depth-first search) charakterisieren. Der Einsatz qualitativer Auswertungsmethoden ermöglicht es, auch Begründungen und Gedankengänge der Probanden bezogen auf konkrete Handlungen aufzuzeigen und dadurch zusätzliche Hinweise auf mögliche Verhaltensunterschiede zu identifizieren. Auf diese Weise können die Autoren bspw. zeigen, dass Probanden der Gruppe mit niedrigem Domänenwissen und hoher Webexpertise sowohl ihr auf die Suche bezogenes als auch ihr domänenbezogenes Wissen zur Bewertung der einzelnen Suchergebnisse heranziehen (ebd., S. 68). Jenkins et al. (ebd., S. 68) illustrieren dies anhand des folgenden Beispiels: Während Probanden dieser Gruppe grundsätzlich Webseiten ihnen bekannter Organisationen bevorzugen, weicht eine Teilnehmerin im Zuge einer ihrer Suchen von dieser Norm ab und wählt stattdessen das auf der Suchergebnisseite als erstes platzierte Suchergebnis aus. Die diesbezügliche Erklärung der Probandin lässt deutlich werden, dass sie sich in diesem Fall auf ihre Sucherfahrung verlässt und somit davon ausgeht, dass die Suchmaschine die Suchergebnisse ihrer Relevanz entsprechend sortiert anzeigt: „This site [the first listed] has a lot of bearing for the search because it is on top.“ (ebd., S. 68)

Zusammengefasst kann festgestellt werden, dass die Beherrschung unterschiedlicher Suchstrategien und die Fähigkeit zur qualitativen Beurteilung von Information den Erfolg der Suche, wie erwartet, positiv beeinflussen. Ihre Anwendung ist jedoch durch die Verfügbarkeit von Domänenwissen auf Seiten der Suchenden limitiert (Hölscher u. Strube, 2000; Jenkins et al., 2003), weshalb im folgenden Abschnitt weitere Erkenntnisse und Forschungsergebnisse hinsichtlich der Bedeutung von Domänenwissen beschrieben werden.

2.1.2.2. Der Einfluss des Domänenwissens auf die Relevanzbewertung

Nachdem im vorangegangenen Abschnitt zunächst die Auswirkungen von Suchexpertise und Domänenwissen auf das Verhalten der Benutzer im Suchprozess beschrieben wurden, soll nun noch etwas ausführlicher geklärt werden, auf welche Art und Weise die Verfügbarkeit von Domänenwissen die Bewertung der Relevanz der Suchergebnisse beeinflusst. Dazu werden im Folgenden exemplarisch die Ergebnisse einiger Studien vorgestellt, deren Fokus hauptsächlich auf den kausalen Erklärungsbeitrag domänenspezifischen Wissens für die Relevanzbeurteilung gerichtet ist. Summa summarum bestätigt sich der Eindruck, den auch die zum Einfluss der Sucherfahrung zitierten Studien vermitteln: Je erfahrener, je routinierter, je geübter eine Person im Umgang

mit dem Thema der Suche ist, desto besser gelingt es ihr auch, geeignete Suchterme auszuwählen, Trefferdarstellungen und Dokumente zielorientiert zu lesen und relevante von irrelevanten Inhalten zu unterscheiden (Wang u. White, 1999; Vakkari u. Hakala, 2000; Dong et al., 2005; Al-Maskari u. Sanderson, 2006). Hat sich im Kontext der Sucherfahrung gezeigt, dass erfahrenere Nutzer ein größeres Repertoire routinemäßiger Verhaltensmuster für die unterschiedlichen Teilprozesse des Suchverhaltens (Anfrageformulierung, Trefferauswahl, Browsingverhalten) besitzen (vgl. Abschn. 2.1.2.1), so lässt sich in Bezug auf die Domänenerfahrung aufzeigen, dass erfahrene Nutzer durch ihren größeren Themenüberblick genauer bestimmen können, wonach sie suchen und infolgedessen auch schneller und effizienter zum Ziel gelangen (Wang u. White, 1999; Kelly u. Cool, 2002).

Im Rahmen der bereits im Kontext der Erwartungshaltung zitierten Studie der Universität Tampere (Vakkari u. Hakala, 2000) wird dargelegt, dass die Verfügbarkeit domänenspezifischen Wissens die Einschätzung der Relevanz neuer Information erleichtert: Je geringer das Vorwissen ist, über das der Suchende verfügt, desto weniger klar formuliert sind seine Relevanzkriterien und desto weniger ist er in der Lage, relevante Inhalte zu identifizieren (ebd., S. 544). Vakkari und Hakala (ebd.) stellen in diesem Zusammenhang z.B. fest, dass die Studienteilnehmer zu Beginn der Suche, wenn ihnen das nötige Themenverständis noch fehlt, vermehrt auf Einführungs- und Überblicksliteratur zurückgreifen (vgl. Abschn. 2.1.1.3). Auch in der Längsschnittstudie von Wang und White (1999) hat sich gezeigt, dass das vorhandene Domänenwissen die Treffsicherheit der Auswahlentscheidung erhöht. Wang und White (ebd., S. 100) begleiten eine Gruppe von 15 Studienteilnehmern bestehend aus 8 Professoren, 6 Doktoranden und einem Masterstudenten über einen Zeitraum von $2{,}5$ Jahren und erfassen ihr Vorgehen bei der Literatursuche im Rahmen eines konkreten Forschungsvorhabens. Es stellt sich heraus, dass Teilnehmer, die mit dem Recherchethema zum Zeitpunkt der Suche vertrauter sind, dazu tendieren, genauer auszuwählen. Sie wählen weniger Dokumente aus, lesen und zitieren jedoch mehr als Teilnehmer, deren Vorwissen zum Zeitpunkt der Suche geringer ist (ebd., S. 103). Die Ergebnisse von Dong et al. (2005) bestätigen den Befund der Treffsicherheit im Kontext einer Relevanzstudie im medizinischen Bereich. In dieser Studie wird u.a. der Einfluss von Domänenwissen auf die Relevanzurteile von 12 Juroren untersucht. Dazu vergleichen die Autoren die Relevanzurteile von 6 Probanden mit und 6 Probanden ohne medizinisches Vorwissen mit den Urteilen eines Mediziners, die als Goldstandard verwendet werden (ebd.). Um die Konsistenz des Goldstandards sicherzustellen, wird dieser mit dem Abstand von einem Jahr von demselben Juror erneut bewertet. Der korrespondierende Kappa-Wert fällt mit $0{,}879$ sehr hoch aus, was auf eine hohe Konsistenz der Bewertung schließen lässt. Bezogen auf den Einfluss des Domänenwissens können Dong et al. (ebd.) zeigen, dass ein höheres Maß an Fachwissen zu größeren Überschneidungen bei der Relevanzbeurteilung durch unterschiedliche Juroren führt. Dieses Ergebnis korreliert mit den Ergebnissen von Al-Maskari und Sanderson (2006), die den Einfluss des Vorwissens bei der Benutzung eines Question-Answering-Systems untersuchen. Die Autoren legen drei verschiedene Erfahrungsniveaus zugrunde: der Benutzer weiß die Antwort; der Benutzer ist mit dem Thema vertraut; der Benutzer hat keinerlei Vorwissen (ebd., S. 135). Auch hier bestätigt sich wieder, dass eine höhere Vertrautheit mit dem Thema der Suche zu einer höheren Accuracy (Verhältnis korrekter Antworten bezogen auf die Gesamtzahl von Fragen) führt (ebd., S. 135). Dieses

Ergebnis ist auch im Kontext der vorliegenden Arbeit relevant, da davon ausgegangen wird, dass die Zufriedenheit der Benutzer mit der erlebten Systemqualität korreliert. Mit Blick auf die durchzuführenden Experimente muss daher das Domänenwissen der Probanden als mögliche Störvariable berücksichtigt werden (vgl. Abschn. 4.2.3).

Darüber hinaus wirkt sich das Domänenwissen auch auf die Effizienz der Relevanzbeurteilung aus. In diesem Zusammenhang wird bspw. von Kelly und Cool (2002) untersucht, inwiefern die Vertrautheit mit dem Thema der Suche die Lesezeit und die Effizienz beeinflusst. Insgesamt nehmen 36 Probanden an dem Experiment teil. Gemessen wird die Zeit, welche die Testpersonen durchschnittlich für das Lesen eines Dokuments benötigten. Effizienz wird als das Verhältnis von als relevant abgespeicherten Dokumenten in Bezug auf die Gesamtzahl aller angesehenen Dokumente definiert (ebd., S. 74 f.). Alle Teilnehmer bearbeiten sechs Suchaufgaben und haben pro Aufgabe je 20 Minuten Zeit. Danach werden sie gebeten, ihr Vorwissen anhand einer 5-stufigen Skala festzulegen. Die Ergebnisse weisen darauf hin, dass sich durch die Verfügbarkeit von Domänenwissen sowohl die Lesezeit verkürzt, als auch die Effizienz erhöht (ebd., S. 75). Die statistische Analyse ergibt eine signifikante Abhängigkeit der Effizienz der Testpersonen von ihrem Vorwissen, während sich für die Lesezeit der Probanden kein signifikanter Unterschied nachweisen lässt (ebd., S. 75). Letzteres könnte der Tatsache geschuldet sein, dass die Länge der betrachteten Dokumente nicht explizit berücksichtigt wird. Auf Grundlage dieser Ergebnisse schlagen ebd. vor, das Nutzerverhalten als Indikator für das Domänenwissen zu verwenden und auf diese Weise auf die individuellen Bedürfnisse des Benutzers zugeschnittene Leistungen anbieten zu können.

In diesem Abschnitt konnte die Bedeutung domänenspezifischen Vorwissens im allgemeinen Kontext der Suche und speziell im Kontext der Relevanzbeurteilung aufgezeigt werden. Weiterhin deutet bereits die kleine Auswahl unterschiedlicher Untersuchungen die methodische Vielfalt an, die in der Praxis hinsichtlich der konkreten Erfassung und Abgrenzung domänenspezifischen Wissens vorliegt. Abschnitt 4.2.3.3 im Methodenteil beschäftigt sich deshalb mit der Frage, wie Sucherfahrungen und Domänenwissen im Rahmen experimenteller Untersuchungen zum Informationssuchverhalten identifiziert und gemessen werden können.

2.2. Motivationale Faktoren: Suchmotivation, Interesse und Einstellung

Neben den kognitiven spielen auch motivationale Faktoren eine nicht unerhebliche Rolle für die Vorhersage von Suchleistungen. Motivation lässt sich ganz allgemein als die Summe der Beweggründe beschreiben, die das Verhalten einer Person bestimmen. So kann die Motivation z.B. das Engagement, mit dem der Suchende vorgeht, die von ihm gewählten Suchstrategien oder die zeitliche Dauer der Suche erklären. In diesem Abschnitt werden deshalb bedeutsame Aspekte der Motivationspsychologie aufgegriffen und ihre Relevanz für informationswissenschaftliche Fragestellungen aufgezeigt. Für den Kontext dieser Arbeit erscheinen insbesondere zwei Bereiche zentral: Dies ist zunächst das Interesse und die innere Beteiligung der Suchenden. Im Mittelpunkt steht dabei die Frage nach dem intrinsischen Interesse. Hierzu werden sowohl theoriegeleitete als auch methodische Erkenntnisse und Studien besprochen. Im zweiten Bereich geht es um die motivationale Bedeutung von Erwartungen. Hier wird das bereits in Abschnitt 2.1.1.1 erwähnte psychologische Konstrukt der Selbstwirksamkeit erneut aufgegriffen und vertieft.

2.2.1. Interesse und Einstellung

Die Motivationsforschung unterscheidet zwischen *intrinsischen* und *extrinsischen* Motivationsquellen. Intrinsisch motivierte Personen handeln aus reiner Freude und eigenem Interesse an der Sache. Ein gutes Beispiel ist das Neugier- und Explorationsverhalten bei Kleinkindern, da dieses Verhalten völlig unabhängig von Belohnungen und anderen externalen Handlungsveranlassungen stattfindet (Krapp u. Ryan, 2002, S. 59). Während also intrinsisch motivierte Verhaltensweisen durch ein starkes persönliches Interesse der Akteure bestimmt sind, werden extrinsisch motivierte Handlungen meist durch Impulse von außen gesteuert (Deci u. Ryan, 1993, S. 225). So lernen extrinsisch motivierte Personen bspw. nicht aus eigenem Interesse am Thema, sondern für gute Noten, die Belohnung der Eltern oder um negative Folgen (z.B. Versetzungsgefährdung) zu vermeiden. In der Motivationsforschung wird weiterhin davon ausgegangen, dass ein wesentlicher Aspekt der Motivation in der Möglichkeit besteht, selbstständig entscheiden und handeln zu können. Die *Theorie der Selbstbestimmung* von Deci und Ryan (ebd., S. 225) geht dabei davon aus, dass „sich motivierte Handlungen nach dem Grad ihrer Selbstbestimmung bzw. nach dem Ausmaß ihrer Kontrolliertheit“ gegeneinander abgrenzen lassen. Dieser Aspekt erscheint in einem kontrollierten Experiment besonders wichtig. Zum einen ist zu überlegen, welchen Einfluss die spezielle Untersuchungssituation auf die Motivation der Testpersonen und damit auf die Qualität und Verlässlichkeit der Ergebnisse hat. Zum anderen stellt sich die Frage, ob auch die Motivation der Testpersonen durch eine entsprechende Gestaltung der Testsituation verbessert werden kann.

Bevor jedoch auf diese Fragen näher eingegangen wird, soll die Theorie der Selbstbestimmung etwas genauer betrachtet werden. Konkret unterscheidet die Selbstbestimmungstheorie zusätzlich zum Zustand der Amotivation fünf Motivationstypen: Die intrinsische Motivation und vier Varianten der extrinsischen Motivation. Diese verschiedenen Ausprägungsstufen extrinsischer Motivation unterscheiden sich primär im Grad ihrer Internalisierung, d.h. der Verinnerlichung gesellschaftlicher Werte, Normen und sozialer Rollen (Krapp u. Ryan, 2002, S. 61). Abbildung 2.3 stellt diesen Zusammenhang grafisch dar. Während eine Person ihr Handeln auf der Stufe der *externalen Regulation* als ausschließlich von außen determiniert betrachtet, nimmt der Grad der Autonomie über Entscheidungen über die Stufen *Introjektion* und *Identifikation*, bis hin zur höchsten Stufe der *Integration* immer weiter zu (Deci u. Ryan, 1993; Ryan u. Deci, 2000; Krapp u. Ryan, 2002). Bei der Introjektion und Identifikation übernimmt die Person Normen und Werte aus der Gesellschaft und richtet ihr Verhalten danach aus. Im Gegensatz zur Introjektion akzeptiert die Person die extern vorgegebenen Ziele in der Stufe der Identifikation als sinnvoll und erstrebenswert und identifiziert sich darüber mit der vor ihr liegenden Aufgabe (Langfeld, 2006, S. 62). Die letzte Stufe schließlich, auf welcher der höchste Grad der Autonomie erlebt wird, ist der intrinsischen Motivation sehr ähnlich. Der Unterschied besteht jedoch darin, dass die verfolgten Ziele ursprünglich nicht die eigenen waren (ebd., S. 62).

Heinström (2006) hebt die Bedeutung der Motivation für den Suchprozess hervor und geht auf unterschiedliche Lern- und Suchstile ein. Im Rahmen einer Längsschnittstudie mit 574 Schülern der Klassen 6 bis 12 verschiedener öffentlicher Schulen in New Jersey, USA kann Heinström (ebd.) zeigen, dass sich unterschiedliche Lerntypen und heterogene motivationale Orientierungen auch auf das Vorgehen bei der Suche auswirken. Zur Erfassung der unterschiedlichen Lernsti-

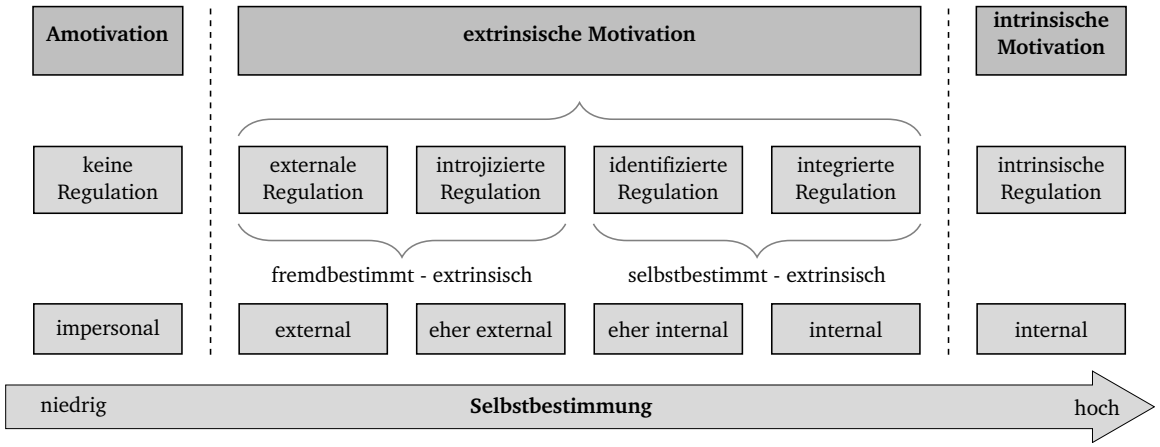


Abb. 2.3.: Motivationstypen der Selbstbestimmungstheorie (nach Ryan und Deci, 2000, S. 72).

le verwendet Heinström (ebd.) eine für Schüler angepasste Version des Approaches and Study Skills Inventory for Students (ASSIST). Auf Basis ihrer Antworten lassen sich die Teilnehmer darüber drei verschiedenen Lernansätzen zuordnen, die sich auch hinsichtlich ihrer motivationalen Grundhaltung unterscheiden: Die erste Gruppe umfasst Teilnehmer, die möglichst wenig arbeiten wollen und einen oberflächlichen Lernansatz (surface approach) wählen. Ihr Lernen ist vollständig extrinsisch motiviert. In Bezug auf die Suche beobachtet Heinström (ebd.) die Bevorzugung leicht zugänglicher Informationsquellen unabhängig von ihrer Qualität. Die zweite Gruppe zeichnet sich durch eine besonders gründliche Vorgehensweise aus (deep approach). Mitglieder dieser Gruppe sind stark intrinsisch motiviert. Ihnen macht das Lernen Spaß, weil sie sich für die angebotenen Themen interessieren. Anhand der Antworten innerhalb dieser Gruppe wird besonders deutlich, welche Faktoren zusammenkommen müssen, um intrinsische Motivation zu fördern: „They found their topics easy because of interest: 'it is pretty easy because I am very interested in the topic', previous knowledge: 'I know some about Italy, so it is not like I'm researching something I know nothing about', or ownership: 'I once visited Egypt so I am really curious to find out more about it'." (ebd.) Insbesondere die letzte Aussage verdeutlicht die wahrgenommene Selbstbestimmtheit, Verantwortung und Bedeutsamkeit der aktuell ausgeübten Tätigkeit. Das Vorgehen bei der Suche betreffend, zeichnet sich diese Gruppe durch ein ähnliches Verhalten wie das von erfahrenen Suchmaschinenbenutzern aus (vgl. Abschn. 2.1.2.1). Ihr intrinsisches Interesse am Thema trägt dazu bei, dass Probanden, die diesen Lernansatz wählen, viele verschiedene Informationsquellen nutzen und diese auch hinsichtlich ihrer Qualität bewerten. Die dritte Gruppe schließlich lässt sich als strategischer Lerntyp charakterisieren (strategic approach). Teilnehmer, die diesen Lernansatz verfolgen, passen sich den jeweiligen Lernzielen an, um möglichst gute Ergebnisse zu erzielen. Im Unterschied zum surface approach verfügt diese Gruppe aber über ein höheres Maß an Selbstverantwortlichkeit und Selbstregulation wie z.B. das folgende Zitat zeigt: „we have this packet to keep all our research in and to keep us organized for our final project" (ebd.). Vom Grad der Internalisierung der Lernaktivität in die eigene Identität befinden sich die Teilnehmer dieser Gruppe vielleicht auf der Stufe der Identifikation. Ihr höheres Maß an innerer Beteiligung bzw. Interesse an der Entscheidung spiegelt sich auch in ihrem Suchverhalten wider, das, wie im Fall des deep approach, durch eine umfassende und kritische Vorgehensweise gekennzeichnet ist (ebd.).

Die Ergebnisse dieser Studie legen nahe, dass die Motivation, mit der ein Nutzer seine Suche beginnt, eine zentrale Rolle bei der Ausführung spielt. Dies gilt sowohl für die Auswahl der Informationsquellen, als auch für ihre anschließende Bewertung. Im Rahmen eines kontrollierten Experiments besteht daher die Gefahr, dass das beobachtete Suchverhalten nicht zwangsläufig mit einer realen Suchsituation übereinstimmt, da die Motivation der Probanden eine andere ist, als in der Realität. So ist z.B. davon auszugehen, dass Probanden in einem IR-Experiment nicht völlig in ihrer Suchtätigkeit aufgehen. Dennoch deutet bspw. der mehrfach geäußerte Kommentar nach Beendigung der im Rahmen dieser Arbeit durchgeführten Experimente, die Teilnahme habe Spaß gemacht, darauf hin, dass ein gewisser Grad der Internalisierung auch in einem kontrollierten Experiment erreicht werden kann. Darüber hinaus lassen sich im Zuge der Studienplanung verschiedene motivationsfördernde Maßnahmen ergreifen, um bei den Probanden das Gefühl der Selbstbestimmung zu unterstützen. Einige von ihnen sollen im Folgenden kurz dargestellt werden.

Wie aus der vorangegangenen Diskussion der Studie von Heinström (2006) deutlich zu erkennen, spielt das Interesse am Thema der Suche eine wesentliche Rolle. Weiter gestützt wird diese These durch eine Untersuchung von Irle (2017), die die Bedeutung des Suchthemas für die Emotionserfahrung während des Suchprozesses hervorhebt. Wichtig ist in diesem Zusammenhang darüber hinaus, dass die Anforderungen der Suchaufgaben zu den Fähigkeiten der Testpersonen passen, damit es weder zu einer Über- noch zu einer Unterforderung kommt. Außerdem sollte vermieden werden, dass die Aufgaben von den Probanden als anstrengend oder belastend empfunden werden. In einigen Studien zum Informationssuchverhalten werden die Testpersonen deshalb gebeten, soweit das Untersuchungsziel dies unter methodischen Aspekten zulässt, ihre eigenen Themen mitzubringen, da auf diese Weise die Chance größer ist, dass die zu bearbeitende Aufgabe als interessant und befriedigend erlebt wird (Su, 1994; Wang u. White, 1999; Vakkari u. Hakala, 2000; Irle, 2017). Edwards und Kelly (2016) führen in diesem Zusammenhang ein interessantes Experiment durch. Um herauszufinden, welchen Einfluss das individuelle Interesse an dem Suchthema auf den Suchprozess hat, lassen die Autoren 40 Probanden im Vorfeld der Untersuchung 8 Testaufgaben in eine nach Interesse sortierte Reihenfolge bringen. Während des Experiments bearbeitet jeder Teilnehmer die beiden von ihm als am meisten und am wenigsten interessant gekennzeichneten Aufgaben. Die anschließende Befragung der Testpersonen hinsichtlich des eigenen Engagements im Rahmen der Aufgabenerledigung ergibt, dass die Probanden bei den interessanten Aufgaben ein signifikant höheres Engagement zeigen. Auch die gemessenen Bearbeitungszeiten lassen einen positiven Zusammenhang mit dem Aufgabeninteresse erkennen. Nichtsdestotrotz können keine signifikanten Unterschiede bezüglich des Suchverhaltens der Testpersonen (z.B. Anzahl gestellter Suchanfragen, Anzahl angesehener Suchergebnisseiten, Anzahl gespeicherter Dokumente) festgestellt werden. Angesichts der Tatsache, dass ein Großteil der Teilnehmer der in dieser Arbeit durchgeführten Experimente aus dem Lehramtsstudium kommt, wird im dritten Experiment auch der Aufgabenfokus auf lehramtsrelevante Themen gelegt (vgl. Abschn. 7.3.4). Auch die Simulation einer konkreten Nutzungssituation kann den Testpersonen helfen, sich besser in die entsprechende Suchsituation hineinzuversetzen (Borlund u. Ingwersen, 1997). Über dieses Rollenspiel und die damit mögliche Identifikation der Probanden mit ihren Aufgaben wird die Suche zu einem emotionalen Erlebnis und gleichzeitig die innere Beteiligung

am Thema der Suche gefördert (intrinsische Motivation). Im ersten Experiment dieser Arbeit sollen sich die Probandinnen deshalb vorstellen, Journalistinnen zu sein und im Rahmen ihrer beruflichen Tätigkeit für einen Artikel zu recherchieren (vgl. Abschn. 5.1). Eine in vielen Studien anzutreffende motivationsfördernde Maßnahme betrifft die Zahlung einer Aufwandsentschädigung für die geopferte Zeit, eine Maßnahme, welche die extrinsische Motivation der Testpersonen beim Bearbeiten der Testaufgaben erhöhen soll (Kelly et al., 2008b; Flavián-Blanco et al., 2011). In den Experimenten dieser Arbeit haben die teilnehmenden Testpersonen die Gelegenheit an der Verlosung von drei Geldpreisen teilzunehmen. Schließlich soll noch eine weitere Maßnahme genannt werden, die in allen drei Experimenten dieser Arbeit sowie im Rahmen einiger anderer Studien Anwendung findet (Chin u. Fu, 2010; Kiseleva et al., 2016; Luo et al., 2017). Um bei den Probanden den Grad der Autonomie über Entscheidungen zu erhöhen, erhalten sie die Möglichkeit, die Aufgaben schon vor Ablauf der vorgegebenen Bearbeitungszeit von zehn Minuten abzuschließen. Dadurch können sich die Testpersonen die Arbeit besser einteilen und haben stärker das Gefühl, selbstbestimmt zu handeln.

Wieder auf die Fragen zu Beginn dieses Abschnitts zurückkommend, kann also einerseits gesagt werden, dass die Motivation der Testpersonen einen nicht zu vernachlässigenden Einflussfaktor hinsichtlich des Suchverhaltens darstellt. Die Möglichkeit, die Motivation im Rahmen einer Kovarianzanalyse statistisch zu kontrollieren, erscheint vor diesem Hintergrund eine geeignete Maßnahme, um die Validität der Ergebnisse zu überprüfen (vgl. Abschn. 4.2.3 u. 4.3.2.3). Andererseits können aber auch untersuchungsdesigntechnische Maßnahmen die motivationale Involviertheit der Testpersonen erhöhen, wenn sie sich z.B. für das zu bearbeitende Thema interessieren.

2.2.2. Suchmotivation durch Erwartung

Ein weiterer Faktor, der die Motivation beeinflusst, ist die persönliche Einschätzung der eigenen Kompetenzen und Fähigkeiten. Diese Selbsteinschätzung, auch Selbstwirksamkeit genannt (vgl. Abschn. 2.1.1.1), spiegelt die Zuversicht einer Person wider, bestimmte Aufgaben aufgrund der eigenen Fähigkeiten bewältigen zu können. Selbstwirksamkeit bezieht sich also nicht auf die für die Erfüllung einer bestimmten Aufgabe, wie z.B. das Auffinden der Öffnungszeiten eines Restaurants im Internet, notwendigen Fähigkeiten, sondern auf das spezifische Selbstvertrauen einer Person, die notwendigen Fähigkeiten für diese Aufgabe zu besitzen. Eine höhere Selbstwirksamkeitserwartung wirkt sich demnach positiv auf die Handlungsabsicht aus, im Internet nach diesen Öffnungszeiten zu suchen. Eine niedrigere Erwartungshaltung hingegen könnte zur Folge haben, dass die betreffende Person andere Informationsquellen wie bspw. das Branchenbuch oder einen Anruf bei dem betreffenden Restaurant in Betracht zieht. Forschungsergebnisse der letzten Jahrzehnte haben den von Bandura (1986) postulierten Zusammenhang zwischen Selbstwirksamkeitserwartungen und Leistungserreichung in den verschiedensten Bereichen bestätigt (Pajares, 1997, S. 1). So stellen Beierlein et al. (2013) dar, dass Selbstwirksamkeitserwartungen in den Bereichen Lernen, Gesundheit, Arbeit, soziale Beziehungen und Politik eine wichtige Rolle spielen. Zusammenfassend zeigen diese Studienergebnisse, dass Personen mit hohen Selbstwirksamkeitserwartungen sich ambitioniertere Ziele setzen, besser mit Stress umgehen können, ein größeres Durchhaltevermögen zeigen und häufig auch effektivere und effizientere Problemlösungsstrategien entwickeln können.

Überträgt man dies auf den Kontext der Suche, so erscheint es plausibel, dass eine hohe Selbstwirksamkeitserwartung in Bezug auf die Internetsuche dazu beiträgt, dass Nutzer sich besser konzentrieren und somit effektivere Suchstrategien anwenden, als Nutzer, die fortwährend am eigenen Erfolg zweifeln. Tatsächlich lässt sich ein ähnlicher Effekt auch im Kontext der Benutzerzufriedenheit nachweisen. In einer Nutzerstudie untersuchen Kelly et al. (2008b), welchen Einfluss Feedback zur Suchleistung der Probanden auf deren Zufriedenheitsurteil ausübt. Dazu bewerten die Teilnehmer nach jeder von drei Suchaufgaben die Leistung des Testsystems. Im Anschluss an diese drei Aufgaben erhalten die Probanden entweder eine negative, realistische, positive oder keine Rückmeldung zu ihrer Suchleistung. In einem finalen Fragebogen werden sie dann gebeten, noch einmal abschließend die Leistung des Systems zu bewerten. Kelly et al. (ebd.) können einen signifikanten Einfluss des Feedbacks auf diese finale Bewertung nachweisen: Teilnehmer, die ein positives Feedback erhielten, korrigierten ihre Einschätzung des Systems nach oben, während Teilnehmer, die ein negatives Feedback erhielten, ihre Einschätzung des Systems nach unten korrigierten. In diesem Sinne hat also die Evaluation der persönlichen Suchleistung direkten Einfluss auf die Nutzerzufriedenheit und es scheint plausibel, dass das Feedback einen ähnlichen Effekt auf die Selbstwirksamkeit der Probanden ausübt. Bandura (1986, S. 394) argumentiert darüber hinaus, dass Personen mit einer hohen Selbstwirksamkeitserwartung davon ausgehen, die vor ihnen liegenden Aufgaben bewältigen zu können und deshalb mit mehr Ausdauer und Beharrlichkeit an diese Aufgaben herangehen. Im Gegensatz dazu erwarten Personen mit niedrigen Selbstwirksamkeitsüberzeugungen eine Niederlage und werden in der Folge auch schneller aufgeben, wenn Erfolge ausbleiben. Überträgt man dies erneut auf den Kontext der Informationssuche, könnte man erwarten, dass eine hohe Selbstwirksamkeitserwartung dazu führt, dass Nutzer bei Schwierigkeiten die Suche nicht einstellen, sondern bspw. alternative Suchstrategien anwenden. Dabei sind Selbstwirksamkeitserwartungen nicht fix, sondern dynamisch und können durch Erfolgserfahrungen und Lernprozesse verändert werden (Beierlein et al., 2013). Eine häufig genannte Forderung seitens der Selbstwirksamkeitsforschung im Bereich der Computer- und Internetnutzung lautet daher, entsprechende Trainings und Support anzubieten, die den Benutzern das Selbstvertrauen vermitteln, das sie brauchen, um selbstständig mit den Systemen arbeiten zu können (Compeau u. Higgins, 1995; Tsai u. Tsai, 2003).

Im Bereich der Computernutzung werden Selbstwirksamkeitserwartungen zur Erklärung von allgemeinen Einstellungen gegenüber Computern (Chen, 1986; Kinzie et al., 1994; Compeau u. Higgins, 1995), als Unterstützungsfaktor für die tatsächliche Nutzung von Computern oder Software-Programmen (Igbaria u. Iivari, 1995; Lindblom et al., 2012) oder als Prädiktor für Computerangst (Sam et al., 2005; Embi, 2007; Simsek, 2011; Cazan et al., 2016) herangezogen. Es gibt wenige Studien, die sich dezidiert mit dem Einfluss von Selbstwirksamkeitserwartungen im IR-Kontext auseinandersetzen. Tsai und Tsai (2003) berichten die Befunde einer Fallstudie mit acht Studenten einer Universität in Taiwan. Wie bereits vermutet, zeigt sich, dass Studenten mit höheren Selbstwirksamkeitserwartungen effektivere Suchstrategien entwickeln und die gefundenen Informationen auch besser einordnen können (ebd.). Dies äußert sich bspw. in einer höheren Bereitschaft, neue Suchterme auszuprobieren, einem kritischeren Umgang mit den erhaltenen Suchergebnissen sowie einem zielgerichteteren Vorgehen bei der Suche (ebd.). Monoi et al. (2005) konzeptualisieren Such-Selbstwirksamkeit in Bezug auf spezifische Such-

kompetenzen und entwickeln ein auf dem Information Literacy Standard (ACRL) aufbauendes Instrument zu deren Erfassung. Angewandt wird das Instrument erstmals im Rahmen eines Online-Seminars zu Suchkompetenzen, wobei die Teilnehmenden den Fragebogen einmal zu Beginn und einmal am Ende des Seminars ausfüllen. Es zeigt sich, dass sich die Kurserfahrungen positiv auf die Selbstwirksamkeitserwartungen der Teilnehmer auswirken: Bei sechs der zwölf Items geben die Befragten nach Beendigung des Kurses einen höheren Wert an, d.h. sie fühlen sich kompetenter und sicherer, was ihre eigenen Suchkompetenzen betrifft (ebd., S. 102). Chiou und Wan (2007) berichten von zwei experimentellen Studien, in denen die Effekte von Aufgabenschwierigkeit und initialer Selbstwirksamkeitserwartung auf die Kompetenzwahrnehmung nach der Sucherfahrung untersucht werden. Die Operationalisierung von Aufgabenschwierigkeit erfolgt in diesen Untersuchungen anhand einer Variation der Bearbeitungszeit (ebd., S. 594). Die Selbstwirksamkeitseinschätzung der Probanden wird auf einer stufenlosen Skala mit Extremen an beiden Enden erfasst und anschließend auf eine 100-stufige Skala standardisiert (ebd., S. 595). Die Ergebnisse der ersten Studie bestätigen die Annahme, dass Erfolgserfahrungen (niedrige Aufgabenschwierigkeit) die Wahrnehmung der Selbstwirksamkeit erhöhen, während Misserfolg (hohe Aufgabenschwierigkeit) eine Senkung der Selbstwirksamkeit zur Folge hat. Es wird zudem beobachtet, dass wiederholte Erfolgserlebnisse einen stetig wachsenden linearen Trend der Kompetenz-Selbsteinschätzung hervorrufen. Wiederholte Misserfolgserlebnisse hingegen tragen zu einem rapideren Abfall der eigenen Kompetenzwahrnehmung bei. Die zweite Studie befasst sich mit dem moderierenden Einfluss initialer Selbstwirksamkeitserwartungen auf die Kompetenzwahrnehmung der Testpersonen nachdem sie entweder eine Suchaufgabe mit niedriger oder hoher Aufgabenschwierigkeit bearbeitet haben. Es zeigt sich, dass der Einfluss der Erfolgserfahrung auf das Selbstwertgefühl bei Probanden mit niedrigeren Ausgangserwartungen ausgeprägter ist als bei Probanden mit einer höheren initialen Kompetenzeinschätzung. Demgegenüber ist der negative Effekt von Misserfolg auf das Selbstwerterleben bei Probanden mit höheren Ausgangserwartungen stärker ausgeprägt (ebd., S. 598). Ohne hier weiter auf die dynamische Verknüpfung von Selbstwirksamkeitserwartungen und Erfahrungen eingehen zu können, lässt sich mit Blick auf die vorliegende Fragestellung dieser Arbeit festhalten, dass ein gleich bleibender Schwierigkeitsgrad der Testaufgaben notwendig ist, um den störenden Einfluss individueller Selbstwirksamkeitserwartungen im Rahmen der Experimente möglichst gering zu halten (vgl. Abschn. 4.1.3.2).

Darüber hinaus sind Selbstwirksamkeitserwartungen abhängig von der zu bewältigenden Aufgabe. Bandura (2006) geht davon aus, dass Personen in unterschiedlichen Situationen und Bereichen unterschiedliche Kompetenzerwartungen haben. Für die Verhaltensvorhersage wäre es demnach am besten, wenn sich die Kompetenzeinschätzung möglichst direkt auf dieses Verhalten beziehen würde. Da dies aus Aufwandsgründen jedoch kaum realisierbar ist, gibt es auch den Ansatz einer kontextübergreifenden Erfassung einer stabilen, allgemeinen Selbstwirksamkeitserwartung. Diesen Ansatz verfolgen auch Beierlein et al. (2013), indem sie eine Kurzskala zur Erfassung der allgemeinen Selbstwirksamkeitserwartungen (ASKU) in deutscher Sprache entwickeln und im Rahmen von drei empirischen Studien validieren. Die Tatsache, dass Selbstwirksamkeitserwartungen auch im Kontext der Suche einen möglichen Erklärungsbeitrag leisten können, macht eine solche Kurzskala prinzipiell auch für die Durchführung experimen-

teller Studien zum Informationssuchverhalten interessant, insbesondere da diese Skala aus nur drei Frageitems besteht und sich somit der zeitliche Aufwand für die Beantwortung der Fragen in einem vertretbaren Rahmen bewegt. Allerdings ist zu beachten, dass die Ergebnisse von Beierlein et al. (2013) zur ASKU-Skala erst nach Abschluss der im Rahmen dieser Arbeit durchgeführten Nutzerstudien veröffentlicht worden sind und somit bei der Fragebogenkonstruktion nicht berücksichtigt werden konnten. Stattdessen kann auf einige von Szajna und Scamell (1993) entwickelte Items zur Erfassung von Benutzererwartungen zurückgegriffen werden, die der Selbstwirksamkeitserwartung recht nahe kommen (vgl. Abschn. 6.3.1).

2.3. Demographische Faktoren: Lebensalter und Geschlecht

Bei den Variablen Alter und Geschlecht handelt es sich um weitere für den Suchprozess relevante Einflussfaktoren, deren Berücksichtigung zu einer genaueren Abbildung des Suchverhaltens führt. Im Folgenden wird zunächst der Einfluss des Lebensalters auf das Suchverhalten analysiert. Es wird erläutert, warum dieser - zunächst naheliegend erscheinende - Einflussfaktor für den Suchprozess möglicherweise weniger bedeutend ist, weil andere Faktoren, wie Such- und Domänenerfahrung, letztlich bspw. längere Lesezeiten aufwiegen. Anschließend wird im nächsten Abschnitt auf Arbeiten zum Einfluss von Geschlechterunterschieden eingegangen. Die diesbezüglichen Befunde für den Suchprozess sind uneinheitlich. Während einige Autoren zwar von Unterschieden berichten, werten andere ihre Studien als Indiz gegen systematische Unterschiede zwischen Frauen und Männern.

2.3.1. Das Alter als Gegenstand informationswissenschaftlicher Forschung

Im Kontext der alternden Gesellschaft gewinnt die Frage, inwieweit das Alter eines Nutzers Einfluss auf seinen Sucherfolg hat, immer größere Bedeutung. Dabei zeigen eine Reihe von Studien, dass sich für ältere Nutzer die Informationssuche im Internet problematischer gestaltet als für jüngere (Tullis, 2007; Zaphiris u. Savtich, 2008; Chevalier et al., 2015). So brauchen ältere Nutzer bspw. häufig mehr Zeit, um den Aufbau einer Webseite zu analysieren, den dargestellten Text zu überfliegen und die für sie relevanten Informationen zu entnehmen. Einige Autoren vermuten, dass im Alter auftretende kognitive Leistungsrückgänge die Hauptquelle der beobachteten Unterschiede darstellen (Pak u. Price, 2008; Dommes et al., 2011; Chevalier et al., 2015; Karanam u. van Oostendorp, 2016). Zur Erklärung wird in diesem Zusammenhang häufig die auf Horn und Cattell (1967) zurückgehende Theorie der *fluiden* und *kristallinen Intelligenz* herangezogen (Pak u. Price, 2008; Dommes et al., 2011; Chevalier et al., 2015; Karanam u. van Oostendorp, 2016). Diese unterscheidet zwei verschiedene Intelligenzfaktoren: Während die fluide Intelligenz durch Gene und biologische Merkmale beeinflusst wird, umfasst die kristalline Intelligenz alle Fähigkeiten, die im Laufe des Lebens erlernt werden. Wesentliche Bestandteile der kristallinen Intelligenz sind z.B. die Allgemeinbildung und das Schulwissen einer Person, aber auch ihr verbales Ausdrucksvermögen sowie ihre soziale Kompetenz. Demgegenüber werden Fähigkeiten wie Flexibilität und Kreativität der fluiden Intelligenz zugeschrieben. Diese Form der Intelligenz tritt insbesondere in Situationen zu Tage, in welchen die Fähigkeit verlangt wird, sich neuen Problemen oder Gegebenheiten anzupassen, ohne auf Lernerfahrungen zurückgreifen zu können. Wie oben bereits angedeutet, wird in der Psychologie davon ausgegangen, dass die fluide Intel-

ligenz im jungen Erwachsenenalter ihren Höhepunkt erreicht und danach sukzessive abnimmt, während die kristalline Intelligenz im Alter weiter zunehmen kann.

Im Kontext des IR spielen beide Intelligenzfaktoren eine Rolle. In der alltäglichen Internetsuche wie auch in fachlichen Domänen benötigt man sowohl das eigene Erfahrungswissen als auch die Fähigkeit, sich auf neue Situationen einzustellen und z.B. verschiedene Suchstrategien auszuprobieren – insbesondere bei schwierigen Aufgaben oder komplexen Themen. Die Studien von Dommes et al. (2011), Chevalier et al. (2015) und Karanam und van Oostendorp (2016) untersuchen in diesem Zusammenhang das Reformulierungsverhalten von älteren und jüngeren Suchmaschinennutzern. Sie variieren dazu systematisch den Schwierigkeitsgrad der gestellten Aufgaben, was dazu beitragen soll, dass die Teilnehmer ihre Suchanfragen häufiger überarbeiten müssen (Dommes et al., 2011, S. 717). Dieser Überarbeitungsprozess erfordert sowohl Fähigkeiten im Bereich der kognitiven Flexibilität (fluide Intelligenz), als auch das nötige Vokabular (kristalline Intelligenz), um die Suchanfragen entsprechend präzisieren zu können (ebd., S. 717). Alle drei Studien kommen dabei zu weitestgehend einheitlichen Befunden. So ist zu beobachten, dass ältere Teilnehmer generell weniger Suchanfragen stellen, seltener reformulieren, weniger Dokumente ansehen und seltener Spezialisierungen und Generalisierungen ihrer Suchanfragen vornehmen (Dommes et al., 2011; Chevalier et al., 2015; Karanam u. van Oostendorp, 2016). Zusätzlich erfassen die Studien die fluide Intelligenz der Teilnehmer und können einen signifikant niedrigeren Wert für die Teilnehmergruppe der älteren Nutzer nachweisen. Diese im Alter auftretenden Leistungsrückgänge der kognitiven Flexibilität werden dann in allen drei Studien als ursächlich für die beobachtete geringere Suchleistung der älteren Teilnehmer angesehen.

Allerdings ist zu beachten, dass die geringe Zahl an Suchanfragen, Reformulierungen und Seitenaufrufen auch mit einer Unerfahrenheit älterer Nutzer im Bereich der Internetsuche zusammenhängen könnte. So berichten bspw. Karanam und van Oostendorp (2016, S. 5722), dass zwar die Computererfahrung der älteren Teilnehmergruppe höher ausfällt, die Erfahrung im Umgang mit Suchmaschinen im Vergleich zu der jüngeren Untersuchungsgruppe jedoch geringer ist. Eine mögliche These wäre also, dass ältere Teilnehmer weniger Suchanfragen formulieren bzw. Reformulierungen vornehmen, weil ihnen die Notwendigkeit solch einer Vorgehensweise nicht bewusst ist. Dies würde im Umkehrschluss auch erklären, warum eine höhere kristalline Intelligenz älteren Nutzern keinen Vorteil verschafft – sie greifen einfach nicht auf Suchstrategien zurück, die bspw. ihr umfangreicheres Vokabular zur Anwendung brächten. Diese Interpretation ist auch in Übereinstimmung mit der Beobachtung von Chevalier et al. (2015, S. 313), dass im Gegensatz zu den jüngeren Teilnehmern die ältere Vergleichsgruppe im Wesentlichen immer dieselbe Suchstrategie verwendet unabhängig von bspw. der Komplexität der Aufgabe. Die erfahrungsbasierten Vorteile älterer Nutzer, wie ein höheres Erfahrungswissen, können somit nicht zur Anwendung kommen. Die beobachtete Korrelation zwischen kognitiver Flexibilität und Suchleistung muss also nicht zwingend auf einen Kausalzusammenhang hinweisen, sondern könnte einfach auf den Unterschied zwischen älteren Nutzern und *digital natives* zurückzuführen sein.

Diese Interpretation im Sinne einer geringeren Erfahrung wird durch eine weitere Gruppe von Studien gestützt, die den Einfluss des Alters auf das Suchverhalten der Probanden mit Hilfe von Eye-Tracking-Experimenten untersucht (Tullis, 2007; Zaphiris u. Savtich, 2008; Hill et al., 2011). In den Studien von Tullis (2007) und Zaphiris und Savtich (2008) kann erneut ein an-

deres Such-/Leseverhalten der älteren Teilnehmer festgestellt werden: Sie benötigen mehr Zeit, um den Inhalt zu erfassen, was im Wesentlichen darauf zurückgeführt werden kann, dass sie im Gegensatz zu den jüngeren Teilnehmern Textinhalte weniger überfliegen, sondern genauer lesen. Tullis (2007) verwenden in diesem Zusammenhang den von Chadwick-Dias et al. (2003) eingeführten Begriff des *cautious clicker*, der zunächst alle Informationen abwägt, bevor er in seiner Suche weitergeht. Wie bereits erwähnt, führen beide Studien die beobachteten Unterschiede nicht primär auf das Alter, sondern auch auf die geringere Sucherfahrung der älteren Probanden zurück. Diese Interpretation wird weiterhin durch die Studie von Hill et al. (2011) gestützt. Anstatt das Suchverhalten junger und älterer Probanden gegenüber zu stellen, vergleichen Hill et al. (ebd.) das Suchverhalten älterer Internetneulinge mit dem Suchverhalten älterer erfahrenerer Internetnutzer. Zwischen diesen beiden Gruppen lassen sich ähnliche Unterschiede wie sonst zwischen älteren und jüngeren Teilnehmern nachweisen. Die Eye-Tracking-Daten der unerfahreneren Nutzer ähneln dabei denen älterer Probanden in anderen Studien: „Previous research has demonstrated that such a profile is associated with ageing [...] but it is also classically associated with confusion and uncertainty." (ebd., S. 1157) Auf der anderen Seite zeichnet sich die erfahrenere Gruppe durch „a much stronger leftwards skew and tighter spread, matching the general profile reported for younger adults" aus (ebd., S. 1157).

Diese Ergebnisse lassen es plausibel erscheinen, dass zumindest im Kontext der Augenbewegungen die Unterschiede zwischen jüngeren und älteren Nutzern zu einem Teil auf die unterschiedlichen Erfahrungshintergründe im Umgang mit Suchmaschinen zurückgeführt werden können. Im Kontext von Nutzerexperimenten erscheint es somit geboten, über das Alter der Probanden hinaus, insbesondere auch die Computer- und Suchmaschinenerfahrung der Probanden zu kontrollieren, bzw. zu erfragen, um diese ggf. im Rahmen einer Kovarianzanalyse aus den Daten herauspartialisieren zu können (vgl. Abschn. 4.3.2.3).

2.3.2. Geschlechterunterschiede in Suchverhalten und Sucherfolg

In fast allen Kontexten der sozial- und kulturwissenschaftlichen Forschung stellt neben dem Alter auch das Geschlecht der Probanden einen möglichen Einflussfaktor dar. Somit ist es nicht verwunderlich, dass diese Frage auch im Rahmen der Informationssuche Beachtung findet. Allerdings kommen die Studien in Bezug auf einen Unterschied im Suchverhalten zwischen weiblichen und männlichen Versuchspersonen nicht unbedingt zu einheitlichen Aussagen.

So finden einige Studien tatsächlich Hinweise auf ein sogenanntes *gender gap*, bei dem Geschlechterunterschiede in Bezug auf den Umgang mit Computern und dem Internet wie auch für das Interesse an und die Einstellung zu computerbezogenen Themen zu beobachten sind (Richter et al., 2001a; Large et al., 2002; Torkzadeh u. van Dyke, 2002; Roy et al., 2003; Roy u. Chi, 2003). Large et al. (2002), Roy et al. (2003) und Roy und Chi (2003) untersuchen bspw. das Suchverhalten und den Sucherfolg von Kindern im Schulkontext. Dabei beobachten Roy et al. (2003), dass beim Verwenden einer Suchmaschine zum Lösen von Aufgaben der Lerneffekt bei Jungen höher ausfällt als bei Mädchen. Dies gilt sowohl für die spezifische Aufgabenstellung als auch für den breiteren thematischen Kontext. Interessanterweise tritt dieser Geschlechterunterschied nicht zu Tage, wenn dieselbe Aufgabe mit Hilfe einer klassischen Bibliothek bearbeitet wird (ebd.). Die Autoren der beiden Studien führen diesen Effekt auf unterschiedliche Suchstrategien der Jungen und Mädchen zurück (Roy et al., 2003; Roy u. Chi, 2003): Während Jungen

zunächst die Suchergebnisse scannen und gegebenenfalls ihre Anfrage ändern, bis sie mit den Suchergebnissen zufrieden sind, scheinen die Mädchen in der Studie eher dazu zu neigen, diese Auswahl erst auf der Ebene der Dokumente zu treffen. In diesem Zusammenhang scheint die Vorauswahl auf Ebene der Suchergebnisse die effizientere Strategie zu sein. Dies ist vergleichbar mit den Unterschieden zwischen erfahrenen und unerfahrenen Webnutzern (vgl. Abschn. 2.1.2), wobei jedoch im Rahmen eines Selbstauskunftsfragebogens kein Unterschied in der Interneterfahrung zwischen den beiden Untersuchungsgruppen festgestellt werden kann (Roy et al., 2003; Roy u. Chi, 2003). Vergleichbare Ergebnisse in Bezug auf ein unterschiedliches Suchverhalten, werden auch von Large et al. (2002) bestätigt.

Demgegenüber können Hargittai und Shafer (2006, S. 441) keine Unterschiede in den Online-Fähigkeiten von Männern und Frauen nachweisen, vielmehr finden sie „[. . .] no statistically significant difference between men's and women's ability to find content on the web once we control for their socioeconomic background and computer and Internet-use experiences. Rather, age, level of education, and experience with the medium are important predictors. Younger users, those with more years of schooling, those with more web-use experience, and users with a computer at work are better at finding content online." Anstatt des Geschlechts sind es also das Vorwissen und die Erfahrung der Testpersonen, die zu einem besseren Abschneiden führen (vgl. Abschn. 2.1.2). Allerdings finden Hargittai und Shafer (ebd.) einen geschlechtsspezifischen Effekt in Bezug auf die Einschätzung der eigenen Fähigkeiten. Frauen schätzen die eigenen Fähigkeiten in Bezug auf die Onlinesuche generell geringer ein. So bezeichnet sich in dieser Studie bspw. keine der weiblichen Teilnehmerinnen als Expertin, während auf der anderen Seite keiner der männlichen Probanden sich selbst als Anfänger einstuft. Obwohl Hargittai und Shafer (ebd.) also keinen signifikanten Leistungsunterschied zwischen den beiden Gruppen nachweisen können, fällt ihre Eigenwahrnehmung unterschiedlich aus. Dies ist im Kontext dieser Arbeit insbesondere im Hinblick auf Erwartungen und Selbstwirksamkeitseffekte interessant (vgl. Abschn. 2.2.2). Darüber hinaus weisen Hargittai und Shafer (ebd., S. 444) darauf hin, dass dieser beobachtete Unterschied ursächlich für die in anderen Studien beobachteten Geschlechterunterschiede sein könnte: „Women's lower self-assessment vis-a-vis web-use ability may affect significantly the extent of their online behavior and the types of uses to which they put the medium." Diese Befunde werden von einer Reihe von Studien bestätigt, die zu dem Ergebnis kommen, dass männliche Nutzer sich zwar nicht nennenswert in ihren Fähigkeiten von weiblichen Nutzern unterscheiden, jedoch oft höhere Selbstwirksamkeitswerte erreichen, während Frauen im Allgemeinen weniger Selbstvertrauen im Technikumgang zeigen (Enochsson, 2005; Whitley, 1997; Torkzadeh u. van Dyke, 2002).

Diese sich scheinbar widersprechenden Beobachtungen in Bezug auf das gender gap werden zumindest teilweise in einer umfangreichen Studie von Richter et al. (2001a) mit 451 Befragten aufgelöst. Die Studie beruht auf dem sog. INCOBI-Instrument, mit dem unterschiedliche Aspekte der Vorerfahrung, des Computerwissens und der Selbstwirksamkeit erfasst werden können. Das Instrument wird im Rahmen der Kontrolle von Störvariablen ausführlicher diskutiert (vgl. Abschn. 4.2.3.3). Für alle erhobenen Variablen lässt sich mit Hilfe von Varianzanalysen ein signifikanter Geschlechtereffekt nachweisen. Genauer wiesen Männer ein „umfassenderes theoretisches und praktisches Computerwissen, eine höhere Vertrautheit mit dem Computer,

mehr Sicherheit im Umgang mit dem Computer und positivere Einstellungen auf als Frauen und nutzten den Computer bereits länger und intensiver.“ (Richter et al., 2001a, S. 71) Es zeigt sich, dass eine Berücksichtigung der Vorerfahrungen und der computerbezogenen Einstellungen diese Geschlechtereffekte zwar tatsächlich reduziert, die beobachteten Unterschiede jedoch nicht vollständig erklären können.

Für alle hier dargestellten Ergebnisse scheint es jedoch darüber hinaus geraten, auch die Aktualität der Studien zu berücksichtigen. So stammt die Studie von Richter et al. (ebd.) bspw. aus dem Jahre 2001 und es ist daher plausibel, dass durch die allgemeine Zunahme der Internetnutzung beginnend mit dem Grundschulalter, eine ähnlich gestaltete Untersuchung heute zu anderen Ergebnissen kommen würde. Diese Vermutung wird durch mehrere aktuelle Studien gestützt, die sich mit der *computer and information literacy (CIL)* beschäftigen (Hohlfeld et al., 2013; Punter et al., 2016; Taylor u. Dalal, 2017). So finden Punter et al. (2016), dass sich das in früheren Studien beobachtete gender gap geschlossen bzw. in einigen Bereichen sogar umgekehrt hat. Der Untersuchung liegen die Daten aus der 2013 durchgeführten International Computer and Information Literacy Study (ICILS) mit 21 teilnehmenden Ländern zugrunde. Es zeigt sich, dass in der Altersgruppe der vierzehnjährigen Schulkinder Mädchen in Bezug auf die Einordnung und Evaluierung von Informationen den Jungen überlegen sind, während hinsichtlich der technischen Anwendungsebene kein Geschlechteruntschied mehr feststellbar ist. In Bezug auf die Selbstwirksamkeit hingegen scheint das gender gap noch nicht ganz überwunden. So berichten Fraillon et al. (2014) auf Grundlage derselben ICILS Studie, dass die Selbstwirksamkeit in Bezug auf fortgeschrittene CIL-Fähigkeiten wie bspw. das Erstellen einer Datenbank bei den männlichen Teilnehmern immer noch höher ausfällt.

Zusammenfassend lässt sich somit festhalten, dass geschlechterspezifische Unterschiede zwar an Bedeutung zu verlieren scheinen, für bestimmte Aspekte des Technikumgangs jedoch nach wie vor nachweisbar sind. Gerade im Hinblick auf die Selbstwirksamkeit scheint es somit vor dem Hintergrund der hier behandelten Forschungsfragen ratsam, einen möglichen Einfluss des Geschlechts der Teilnehmer zu überprüfen bzw. zu kontrollieren. Während im bisherigen Teil des Kapitels die primär von der Person abhängenden individuellen Einflussfaktoren Beachtung finden, wird im folgenden Unterkapitel eine situative Perspektive eingenommen, d.h. es wird untersucht, wie sich die Handlungen und insbesondere die Handlungsentscheidungen der Probanden zu den situativen Bedingungen während der Suche verhalten.

2.4. Situative Faktoren: Aufgabenschwierigkeit und Aufgabenkomplexität

Neben individuellen Einflüssen muss bei IIR-Studien auch das subjektive Erleben der Suchsituation sowie die sich daraus ergebende Dynamik berücksichtigt werden. Im Vergleich zu den bisher besprochenen Personenfaktoren sind derartige Umweltfaktoren in einer experimentellen Untersuchung einfacher zu kontrollieren. Eng mit dem Forschungsdesign verknüpfte Einflussfaktoren, wie Testsystem, Testkorpus und Operationalisierung von Systemgüte und Erwartungshaltung, werden eingehender im Methodenteil in Kapitel 4 behandelt. Situative Faktoren, die eher dynamische Aspekte der Perzeption von Suchergebnissen beeinflussen, werden dagegen im Rahmen der Zielfaktoren Relevanzwahrnehmung und Nutzerzufriedenheit in den Abschnitten 3.1.3 bzw. 3.3.2.4 besprochen. Der folgende Abschnitt beschäftigt sich hingegen eingehender mit Studien

zur Schwierigkeit und Komplexität von Suchaufgaben, die Wildemuth et al. (2014, S. 1120) zufolge die in der Literatur am häufigsten genannten Differenzierungsprinzipien darstellen.

In einer Metastudie werten Wildemuth et al. (ebd.) über 100 IR-Experimente aus. Ihre Untersuchung beschäftigt sich mit der Fragestellung, welche Dimensionen die beiden Konstrukte Aufgabenkomplexität und Aufgabenschwierigkeit ausmachen und welche Bedeutung die einzelnen Dimensionen innerhalb des Konstrukts haben. Während die Komplexität einer Aufgabe häufig anhand von objektiven Kriterien wie bspw. der Anzahl zu erledigender Unteraufgaben, der Anzahl unterschiedlicher Bedeutungsaspekte (Facetten) oder der Anzahl auszuwertender Quellen gemessen wird, hängt die Schwierigkeit einer Aufgabe stärker von der Aufgabenwahrnehmung und dem Vorwissen der einzelnen Testpersonen ab (ebd.). Dieser Vergleich lässt folgende Rückschlüsse zu. Zum einen ergibt sich für die Konstruktion von Testaufgaben, dass die Komplexität der Aufgaben in einem IIR-Experiment leichter zu beherrschen ist, da diese bspw. über die konstante Vorgabe der von Testpersonen zu erbringenden Leistung umgrenzt werden kann (vgl. Abschn. 4.1.3.2). So verwendet der überwiegende Teil interaktiver IR-Studien klassische *ad-hoc* Suchaufgaben, bei welchen die Testpersonen so viele relevante Dokumente wie möglich in einer bestimmten Zeitspanne finden sollen. In anderen Studien hingegen wird explizit verlangt, Dokumente mit unterschiedlichen Inhaltsaspekten ausfindig zu machen (*instance recall*) (Hersh et al., 2000; Allan et al., 2005; Al-Maskari et al., 2006). Während erstere eher zu einem explorativen Verhalten einladen, bei dem lediglich die allgemeine Relevanz der Dokumente erfasst werden muss, erfordert eine Instance-Recall-Aufgabe ein umfassenderes Verständnis des Suchthemas, da in diesem Fall dem speziellen Informationsgehalt der Dokumente eine größere Aufmerksamkeit zuteil wird. Zum anderen macht dieser Vergleich deutlich, dass das subjektive Erleben der Suchsituation eine entscheidende Rolle in Bezug auf den Einfluss der gewählten Aufgaben spielt. Vor dem Hintergrund, dass es hier vorrangig um das individuelle Such- und Bewertungsverhalten der Teilnehmer interaktiver Retrievaltests geht, erscheint demnach insbesondere der Einfluss der Aufgabenschwierigkeit von Interesse. Wildemuth et al. (2014, S. 1129) identifizieren in diesem Zusammenhang folgende vier Aspekte, die geeignete Prädiktoren für die Aufgabenschwierigkeit darstellen: die erreichte Suchleistung, die Anzahl der relevanten Dokumente innerhalb der Kollektion, eine übereinstimmende Terminologie in Aufgabenbeschreibung und Suchergebnissen sowie der wahrgenommene Schwierigkeitsgrad durch den Suchenden. Bevor jedoch im Folgenden näher auf diese unterschiedlichen Aspekte der Aufgabenschwierigkeit eingegangen wird, soll zunächst noch eine Studie vorgestellt werden, die den Komlexitätsgrad von Suchaufgaben aus Sicht der Benutzer analysiert.

Li et al. (2011) führen eine Fragebogenstudie mit 168 Teilnehmern durch. Sie erfassen den Komplexitätsgrad von sechs simulierten Arbeitsaufgaben (vgl. Abschn. 4.1.3.2), indem sie objektive Kriterien der Aufgabenkomplexität mit der subjektiven Wahrnehmung der Befragten vergleichen. Konkret identifizieren Li et al. (ebd.) sieben objektive Maße, um die Komplexität der Aufgaben zu analysieren: die Anzahl der Schlüsselbegriffe in der Aufgabenbeschreibung, die Anzahl zu erledigender Unteraufgaben, die Anzahl verwendeter Terminologien, die Anzahl unterschiedlicher Sprachen in den Suchergebnissen, die Anzahl schwer verständlicher Begriffe in der Aufgabenbeschreibung, die Komplexität der Syntax sowie die Anzahl betroffener Suchdomänen. Die Teilnehmer der Studie bewerten diese Kriterien auf einer 5-stufigen Skala. In Bezug

auf die verwendeten objektiven Maße zeigt sich, dass die Anzahl schwer verständlicher Begriffe, die Anzahl unterschiedlicher Sprachen sowie die Anzahl betroffener Domänen am stärkten mit der Aufgabenkomplexität korrelieren. Darüber hinaus ergibt die Analyse der Fragebögen, dass die Wahrnehmung schwer verständlicher Begriffe in der Aufgabenbeschreibung am stärksten mit der objektiven Komplexität der Aufgaben korreliert. Diese Ergebnisse lassen Li et al. (2011, S. 6) folgern, „[. . .] that task complexity reflects users' capability of understanding a task and its outcomes." Die im Folgenden vorgestellten Studien greifen jeweils unterschiedliche Aspekte der Aufgabenschwierigkeit heraus, um ihren Einfluss zu analysieren.

Smith (2008) erfasst die Aufgabenschwierigkeit in Abhängigkeit der zurückgegebenen relevanten Dokumente und des Vorwissens der Teilnehmer (vgl. Smith und Kantor (2008), Abschn. 3.2.1). Dazu werden die von allen Teilnehmern bearbeiteten Aufgaben mit Hilfe einer im Nachhinein durchgeführten Clusteranalyse entweder als leichtes oder schwieriges Thema identifiziert, je nachdem wie viele relevante Dokumente das Suchsystem zurückliefert und wie vertraut die Teilnehmer mit dem Thema sind. In einem zweiten Schritt wird analysiert, ob die Klassifikation als leichtes oder schwieriges Thema einen signifikanten Einfluss auf das Verhalten der Testteilnehmer hat. Es zeigt sich, dass bei den als schwieriger eingestuften Aufgaben die Frequenz der gestellten Suchanfragen signifikant höher ausfällt als für die leichteren Aufgaben. Smith (2008) kann also einen direkten Einfluss der Aufgabenschwierigkeit auf das Verhalten der Probanden nachweisen. Dieser Zusammenhang wird auch in einer ähnlich angelegten Studie von Chang et al. (2016) bestätigt.

Cox und Fisher (2004) befassen sich mit der wahrgenommenen Aufgabenschwierigkeit und ihrem Einfluss auf die Zufriedenheit der Testteilnehmer. Zu diesem Zweck präsentieren sie den Teilnehmern vier Aufgaben und lassen sie den Schwierigkeitsgrad einschätzen (vgl. Abschn. 2.1.1.3). Die anschließend präsentierten Ergebnislisten sind so konstruiert, dass sie die Erwartungshaltung der Teilnehmer entweder bestätigen oder positiv bzw. negativ enttäuschen. Es zeigt sich, dass der Unterschied zwischen erwarteter Aufgabenschwierigkeit und Qualität der dargebotenen Ergebnisliste stark mit der Nutzerzufriedenheit korreliert. Eng damit verbunden sind die Ergebnisse von Smucker und Jethani (2010b), die umgekehrt zeigen können, dass eine höhere Suchmaschinenqualität zu einer als geringer wahrgenommenen Aufgabenschwierigkeit führt.

Daneben gibt es noch weitere interessante Ansätze zur Kontrolle der Aufgabenschwiergkeit. So ermitteln Chiou und Wan (2007) in Pretests eine realistische Bearbeitungszeit für die von ihnen gestellten Aufgaben. Zur Operationalisierung der Aufgabenschwierigkeit wird den Teilnehmern nun entweder mehr oder weniger Zeit zur Lösung der Aufgaben zugestanden. In diesem Studiendesign untersuchen Chiou und Wan (ebd.) den Einfluss negativer und positiver Erfahrungen (schwere bzw. leichte Aufgaben) auf die Selbstwirksamkeit der Testpersonen (vgl. Abschn. 2.2.2). Wie erwartet erhöhen positive Erfahrungen die Selbstwirksamkeit, während negative Erfahrungen die Selbstwirksamkeit reduzieren. Es zeigt sich jedoch, dass der Effekt negativer Erfahrungen im Vergleich stärker ausfällt (ebd.). In einem ähnlichen Kontext analysieren Liu und Wei (2016) den Unterschied in den angewendeten Suchstrategien bei Suchen mit und ohne Zeitlimit. Sie können zeigen, dass ein Zeitlimit tatsächlich zu einem anderen Nutzerverhalten führt, bei dem die Teilnehmer selektiver bei der Auswahl von Dokumenten aus den Ergebnislisten vorgehen. Die im Kontext von Abschnitt 2.3.1 diskutierten Studien zum Einfluss des Alters auf das Such-

verhalten hingegen greifen auf die dritte Dimension von Wildemuth et al. (2014) zurück, um die Aufgabenschwierigkeit zu variieren (Dommes et al., 2011; Chevalier et al., 2015; Karanam u. van Oostendorp, 2016). Um einen höheren Schwierigkeitsgrad zu generieren, stellen sie sicher, dass die in der Aufgabenbeschreibung verwendete Terminologie nicht zu einem Sucherfolg führt, sondern auf alternative Suchterme zurückgegriffen werden muss. Im Extremfall werden sogar Aufgaben gestellt, die nicht mit einer Webanfrage gelöst werden können (Dommes et al., 2011; Chevalier et al., 2015).

Die in diesem Abschnitt vorgestellten Studien zeigen, dass Aufgabenkomplexität und Schwierigkeitsgrad tatsächlich Einfluss auf das Nutzerverhalten sowie das Sucherlebnis nehmen können. Dabei wird deutlich, dass es vielfältige Ursachen für einen erhöhten Schwierigkeitsgrad, wie Vorwissen, Zeitdruck sowie erwartete Schwierigkeit, gibt. Dies lenkt den Blick auf die grundlegende Bedeutung der Auswahl der Suchaufgaben für das Design einer experimentellen Studie zum Informationssuchverhalten. Neben einer ausreichenden Bearbeitungszeit sollte insbesondere darauf geachtet werden, dass die Formulierung der Aufgaben nicht zu Problemen bei der Suche führt, da ansonsten das Vorwissen der Testpersonen einen größeren Einfluss auf die Suchleistung ausübt. Weitere Aspekte der Aufgabenauswahl werden in Abschnitt 4.1.3.2 im Rahmen des Methodenkapitels behandelt.

2.5. Fazit: Einflussfaktoren auf die Wahrnehmung von Suchergebnissen

In diesem Kapitel wird herausgearbeitet, wie situative Kontexte aber auch individuelle Voraussetzungen der Testpersonen Einfluss auf das zu beobachtende Suchverhalten nehmen. Sowohl die Anzahl als auch die Breite der vorgestellten Einflussfaktoren illustrieren, dass der traditionelle systemorientierte Ansatz nicht ausreicht, um das Suchverhalten von realen Nutzern in seiner Gesamtheit zu erfassen. Gleichzeitig wird deutlich, dass es nicht möglich ist, jeden Effekt auf einen einzelnen Einflussfaktor zurückzuführen. Vielmehr entsteht das beobachtete Verhalten aus einem Zusammenspiel verschiedener Faktoren. Bei der Konzeption eines Untersuchungsdesigns ist somit darauf zu achten, möglichst viele Einflussgrößen experimentell zu kontrollieren, damit die Wirkung der variierten Faktoren klar zu Tage tritt. Darüber hinaus wird deutlich, dass insbesondere die persönlichen Hintergründe, wie bspw. Motivation, Vorerfahrung und Erwartung, der Testpersonen in die Betrachtung des Suchprozesses mit einbezogen werden müssen, da sie unmittelbar Einfluss auf die angewendeten Suchstrategien und Relevanzkriterien ausüben können. Im Umkehrschluss bedeutet dies auch, dass zur Erzielung eines möglichst realitätsnahen Sucherlebnisses der entsprechende Kontext, d.h. das Informationsbedürfnis und die Motivation im Untersuchungsdesign bereitgestellt werden muss. Hier kann auf die Darstellungen und Konzeptionen simulierter Arbeitsaufgaben zurückgegriffen werden.

Weiterhin zeigt die Auseinandersetzung mit den besprochenen Einflussfaktoren, dass eine einseitige Betrachtung allein des Verhaltens oder der Äußerungen der Probanden nicht zielführend ist, wie sich z.B. am eingangs erwähnten Beispiel der boom box belegen lässt. Vielmehr weisen gerade Diskrepanzen zwischen Selbstwahrnehmung und tatsächlichem Verhalten auf interessante Effekte hin. Ein Beispiel aus dem IR-Kontext stellt bspw. die geschlechterabhängige Unter- bzw. Überschätzung der eigenen Suchleistung dar. Erst eine Kombination beider Informationsquellen kann also ein verlässliches und umfassendes Bild des Informationssuchprozesses liefern.

Gleichwohl die Erwartungshaltung der Probanden im IR-Kontext bisher nur vereinzelt Beachtung gefunden hat, kann ihre grundsätzliche Bedeutung für den Suchprozess herausgearbeitet werden. Neben einer vertieften Diskussion des Entstehungsprozesses von Erwartungen kann insbesondere nachgewiesen werden, dass es möglich ist, im Rahmen von IR-Experimenten bei den Testpersonen eine bestimmte Erwartungshaltung zu evozieren. Somit ist eine der entscheidenden Voraussetzungen für die im praktischen Teil dieser Arbeit durchgeführten Benutzerstudien erfüllt.

3. Bewertung von Suchergebnissen: Relevanz, Sucherfolg und Zufriedenheit

Der Paradimgenwechsel von system- zu benutzerorientierter IR-Evaluierung erfordert die Operationalisierung des individuellen Sucherfolgs. Die im vorangegangenen Kapitel in ihrer Vielschichtigkeit vorgestellten Einflussgrößen wie Erwartung, Vorerfahrung, Motivation und demographische sowie situative Faktoren machen für ein solches Konstrukt einen mehrdimensionalen Ansatz, der den gesamten Suchprozess mit in die Betrachtung einbezieht, plausibel. Aus diesem Grund finden neben der Relevanzwahrnehmung als Grundlage der betrachteten Zielfaktoren sowohl die wahrgenommene Effektivität aus Sicht der Probanden, d.h. ihre Zufriedenheit, als auch Verhaltens- bzw. Leistungsmaße zur objektiven Bestimmung des Nutzererfolgs Berücksichtigung. Ein solches Vorgehen wird auch von Frøkjær et al. (2000, S. 345) im allgemeinen Kontext von Usability vertreten: „Unless domain specific studies suggest otherwise, effectiveness, efficiency, and satisfaction should be considered independent aspects of usability and all be included in usability testing."

Der weitere Aufbau dieses Kapitels ist wie folgt gegliedert. Nach einer Diskussion der Relevanzwahrnehmung in Abschnitt 3.1 stellt Abschnitt 3.2 den Stand der Forschung bezüglich der Übertragbarkeit und Wahrnehmung qualitätsbezogener Systemunterschiede aus Nutzersicht dar. Daran anschließend werden in Abschnitt 3.3 die Grundlagen der Nutzerzufriedenheit im IR-Kontext erläutert.

3.1. Die Beurteilung der Relevanz von Informationsobjekten

Der Beurteilung von Relevanz kommt im Kontext der Evaluierung von IR-Systemen eine Schlüsselrolle zu. Tatsächlich bildet die Unterscheidung zwischen relevanten und irrelevanten Dokumenten die Grundlage für die meisten Effektivitätsmaße. Im Gegensatz zum systemzentrierten Ansatz, der Relevanz als objektive Größe begreift, muss dieses Konstrukt im Kontext interaktiver IR-Experimente vom Nutzer her gedacht werden. Dieser Übergang zu einem situativen Relevanzverständnis stellt einen der Hauptunterschiede zwischen system- und benutzerorientierter IR-Evaluierung dar. Allerdings geht diese Erweiterung des Begriffs mit einer erhöhten Komplexität einher, da ähnlich wie im vorangegangenen Kapitel situativen und individuellen Einflussfaktoren Rechnung getragen werden muss. Im Folgenden wird deshalb zunächst auf verschiedene Definitionen von Relevanz eingegangen, um den in dieser Arbeit verwendeten Relevanzbegriff näher zu bestimmen. Im Anschluss daran folgt ein kurzer Überblick über Studien zu Relevanzkriterien und zum Entscheidungsverhalten von Suchmaschinennutzern (vgl. Abschn. 3.1.2). Im letzten Unterabschnitt werden darüber hinaus verschiedene Überlegungen zur Wahrnehmung von Relevanz, insbesondere zu Verhaltensänderungen bei der Relevanzbeurteilung, aufgezeigt und erläutert (vgl. Abschn. 3.1.3).

3.1.1. Das Konstrukt der situativen Relevanz

In der Literatur existiert eine Vielzahl von Ansätzen und Theorien zur Beschreibung von Relevanz (Schamber, 1994; Borlund, 2003a; Saracevic, 2007a; Saracevic, 2007b). In ihrer einfachsten Form lässt sich Relevanz als der von einem *Benutzer* wahrgenommene Grad der Übereinstimmung zwischen einem *Informationsbedürfnis* und einem *Dokument* definieren (Saracevic, 1975, S. 328).

Verschiedene Autoren betonen die Multidimensionalität des Konstrukts der Relevanz, die durch einen unidimensionalen Ansatz nur ungenügend wiedergegeben werden kann. So erachtet Borlund (2003a) ein multidimensionales und dynamisches Konzept für zweckmäßiger, um interaktives Retrieval abzubilden. Dieser Meinung ist auch Anderson (2005), die zudem darauf hinweist, dass die Bewertung von Relevanz als Prozess verstanden werden muss: „From the perspective of searchers engaged in task-based information-seeking, relevance is more than the selection or rejection of information. It is a multi-level phenomenon communicated through the absence as well as the presence of connections that researchers recognize at the time.“ Auch Harter (1992, S. 603) plädiert für ein multidimensionales Verständnis von Relevanz, indem er eine rein inhaltliche Betrachtung als zu eng gefasst zurückweist: „Relevance, in its everyday sense, is much more complex than simple topicality.“ Die Erscheinungsformen von Relevanz (manifestations of relevance) nach Saracevic (1996, S. 214) stellen eine umfassende Beschreibung des multidimensionalen Charakters der Relevanz dar, die im Folgenden kurz erläutert werden (vgl. Borlund, 2003a; Toms et al., 2005):

Systemrelevanz (system/algorithmic relevance) – Die Systemrelevanz drückt die Ähnlichkeit zwischen Suchanfrage und Suchergebnis aus. Diese Form der Relevanz hängt von der jeweiligen Implementierung des Suchalgorithmus ab und bezieht sich auf den objektiven, nutzerunabhängigen Nutzen eines Dokuments. Laut Borlund (2003a, S. 914 f.) handelt es sich hierbei um die wohl gebräuchlichste und eindeutigste Definition von Relevanz, die auch bei der traditionellen Evaluierung nach dem Cranfield-Paradigma im Mittelpunkt steht.

Thematische Relevanz (topical/subject relevance) – Diese Form der Relevanz bezieht sich auf die semantische Übereinstimmung der Suchergebnisse mit dem Thema der Suche. Zur besseren Unterscheidung zwischen Systemrelevanz und thematischer Relevanz spricht Borlund (ebd., S. 915) im zweiten Fall von *intellektueller, thematischer Relevanz* (intellectual topicality) und legt damit ein stärkeres Gewicht auf die subjektive Bewertung der Relevanz durch Benutzer oder Juroren.

Kognitive Relevanz (cognitive relevance/pertinence) – Die kognitive Relevanz repräsentiert die intellektuelle Beziehung zwischen dem individuellen Informationsbedürfnis des jeweiligen Benutzers und dem wahrgenommenen Nutzen der soeben betrachteten Informationsobjekte. Die Einschätzung der Relevanz basiert in diesem Fall auf den persönlichen Erfahrungen der Benutzer und spiegelt ihre persönliche Meinung zu einem bestimmten Zeitpunkt wider. Auch wenn thematische und kognitive Relevanz auf den ersten Blick recht ähnlich erscheinen mögen, gibt es doch entscheidende Unterschiede. So können neben der zentralen Fragestellung der Dokumente auch andere Aspekte wie Informationsgehalt, Neuheitswert oder Informationsqualität zur Beurteilung der kognitiven Relevanz herangezogen werden (Saracevic, 1996). Borlund (2003a, S. 915) zufolge unterscheidet sich dieser Relevanztyp

vor allem durch die dynamische Sichtweise der Informationsbedürfnisse.

Situative Relevanz (situational relevance/utility) – Bei dieser Form der Relevanz kommt der situative Kontext als weitere Komponente hinzu. Die situative Relevanz wird somit als der kontextspezifische Nutzen eines Dokuments verstanden, der sich aus der Betrachtung von Situation, Aufgabe und Dokumenten ergibt. Da auch dieser Relevanztyp zudem überaus dynamisch ist, lässt sich laut Borlund (ebd., S. 915), aus Beobachtersicht häufig nicht mit Bestimmtheit sagen, ob es sich im konkreten Fall um kognitive oder situative Relevanz handelt. Saracevic (1996) nennt folgende Indikatoren für situative Relevanz: den Nutzen für die Entscheidungsfindung, die Sachdienlichkeit der Dokumente zur Lösung des Problems sowie die Reduktion von Unsicherheit.

Motivationale Relevanz (motivational/affective relevance) – Der Bereich der motivationalen Relevanz umfasst Beziehungen zwischen den Absichten, Zielen und Motiven des jeweiligen Benutzers und dem Ergebnis der Suche. Zufriedenheit, Erfolg und Aufgabenerfüllung sind mögliche Kriterien für eine Evaluierung (ebd.). Laut Borlund (2003a, S. 915) handelt es sich hierbei nicht um eine eigenständige Relevanzdefinition. Vielmehr seien in allen von Saracevic (1996) beschriebenen subjektiven Formen von Relevanz motivationale und affektive Strategien enthalten: „[. . .] the ‚drive‘ to want information is not an independent, specific type of relevance, but an inherent characteristic of relevance behavior in general.“ (Borlund, 2003a, S. 915)

Im Vergleich zu Kapitel 2 wird deutlich, dass viele der hier genannten Aspekte auch schon im Kontext der individuellen und situativen Einflussfaktoren diskutiert worden sind. In Anlehnung an Wilson (1973), Schamber et al. (1990), Harter (1992), Borlund und Ingwersen (1997) und Borlund (2000) wird im Rahmen dieser Arbeit deshalb von einer situationsbedingten Definition der Relevanz aus Benutzerperspektive ausgegangen. Anderson (2005) formuliert die Notwendigkeit, den jeweiligen Kontext zu berücksichtigen folgendermaßen: „Notions of the concept of relevance can thus appear contradictory or inconsistent to an observer who is not aware of the context in which relevance assessment is made. Relevance cannot be examined in isolation from the particular situation in which information is pursued, evaluated and utilised.“ Wilson (1973, S. 460) betont darüber hinaus die Individualität des einzelnen Nutzers: „Situational relevance is relevance to a particular individual’s situation - but to the situation as he sees it, not as others see it or as it ‚really‘ is .“

Gemäß dieser Definitionen ist die Zuschreibung von Relevanz subjektiv und situationsspezifisch, weil sich Nutzer im Hinblick auf Erfahrungen, Erwartungen, Bedürfnisse, Vorlieben, Zielsetzungen usw. voneinander unterscheiden (vgl. Kap. 2). Diese Einflussfaktoren werden bei der situativen Relevanz als wesentlicher Bestandteil des Beurteilungsprozesses verstanden, sodass die Frage, was als relevant angesehen wird und was nicht, einem ständigen Wandel und Anpassungsprozess unterworfen ist. Um in benutzerorientierten, interaktiven Suchprozessen relevant zu sein, müssen Informationsobjekte Veränderungen in den kognitiven Strukturen der Nutzer bewirken. Dies kann geschehen, indem neues Wissen gewonnen und mit bereits bekanntem verbunden oder altes Wissen korrigiert, ergänzt oder verworfen wird: „A phenomenon that is relevant to ‚a matter at hand‘ changes the matter in some way; it adds information, or decreases

information, offers a new perspective, or causes other kinds of cognitive change.“ (Harter, 1992, S. 603). Schamber et al. (1990, S. 774) fassen den situativen Ansatz anhand von drei Prämissen folgendermaßen zusammen:

1. Relevance is a multidimensional cognitive concept whose meaning is largely dependent on users' perceptions of information and their own information need situations.
2. Relevance is a dynamic concept that depends on users' judgments of the quality of the relationship between information and information need at a certain point in time.
3. Relevance is a complex but systematic and measurable concept if approached conceptually and operationally from the user's perspective.

Die folgenden zwei Abschnitte gehen genauer auf die ersten beiden dieser drei Prämissen ein und erläutern ihre Implikationen anhand ausgewählter Studien. Der Aspekt der Messung von Relevanz im IIR-Kontext wird hingegen in Abschnitt 4.2.2.1 im Rahmen der methodischen Vorgehensweise diskutiert.

3.1.2. Kriterien zur Relevanzbeurteilung

Es wurde bereits darauf hingewiesen, dass Relevanz ein multidimensionales Konzept darstellt, das auf der situativen Wahrnehmung des einzelnen Nutzers beruht. In der konkreten Nutzungssituation müssen diese Komponenten und Einflussfaktoren jedoch auch erkannt und identifiziert werden. Der folgende Abschnitt stellt einige konkrete Ansätze dazu vor.

Schamber (1994, S. 11) bspw. veröffentlicht eine aus der Literatur zusammengestellte Übersicht von 80 verschiedenen Faktoren, die bei der Bewertung der Relevanz von Suchergebnissen in Betracht gezogen werden können. Dazu gehören bestimmte Kennzeichen der Informationsobjekte wie Autor, Publikationsorgan, Aktualität aber auch situationsspezifische Merkmale wie die Gesamtheit der verfügbaren Dokumente, Usability, Dringlichkeit sowie personenabhängige Kriterien wie Neuheitswert, Verständlichkeit und Qualität der Dokumente.

Eine Reihe von Studien befragt Nutzer mit realen Informationsbedürfnissen hinsichtlich ihrer Auswahlkriterien für relevante Suchergebnisse (Park, 1993; Fidel u. Crandall, 1997; Fitzgerald u. Galloway, 2001; Maglaughlin u. Sonnenwald, 2002). Alle Autoren kommen insgesamt zu ähnlichen Ergebnissen, was für eine breite Gültigkeit der erhobenen Kriterien spricht. Park (1993) unterscheidet bspw. benutzer- und dokumentbezogene Faktoren. Die benutzerbezogenen Faktoren werden wiederum in drei Kontexte untergliedert: den internen Kontext, der Kenntnisse, Fähigkeiten und Erfahrungen des betreffenden Nutzers umfasst, den externen Kontext, welcher die gegenwärtige Situation des Nutzers beschreibt und den Problemkontext, der die gegebene Problem- oder Aufgabenstellung beinhaltet. Die dokumentbezogenen Faktoren reichen von biographischen Merkmalen des Autors bis hin zu charakteristischen Merkmalen des Publikationsorgans. Ihre Wahrnehmung durch den Nutzer wird laut Park (ebd.) in starkem Maß durch die drei Kontexte beeinflusst.

Barry und Schamber (1998) vergleichen die Ergebnisse zwei voneinander unabhängiger Studien (Barry, 1994; Schamber, 1991), um Ähnlichkeiten und Unterschiede in den darin genannten Relevanzkriterien zu erkennen. Auch sie identifizieren zehn übereinstimmende Kriterien, was Barry und Schamber (1998, S. 219) von der Existenz eines begrenzten Kriterienkanons ausgehen

lässt: „The results of the studies, taken together, provide evidence that a finite range of criteria exists and that these criteria are applied consistently across types of information users, problem situations, and source environments. “

Ausgehend von Saracevic's *manifestations of relevance* (Abschn. 3.1.1) entwickeln Toms et al. (2005) Relevanzmaße, die auf die Operationalisierung der von Saracevic (1996) aufgestellten Relevanztypen abzielen. Als Untersuchungsgegenstand dienen sowohl die Interaktionsprotokolle der im Test durchgeführten Websuchen als auch die Rückmeldungen der Testpersonen in Fragebögen und Interviews (Toms et al., 2005). Mittels Faktorenanalyse lassen sich die so ermittelten elf Relevanzmaße zu den drei Gruppen Benutzer, Aufgabe und System zusammenfassen.

Howard (1994) setzt die von Kelly (1955) vorgelegte *Repertory-Grid-Methode* ein, um individuelle Konstrukte über die Auswahl relevanter Dokumente zu erfassen. Fünf bis sieben von den Teilnehmern der Studie selbst ausgewählte Dokumente dienen während der Befragung als Datenbasis. Durch die Vorlage verschiedener Dokumentkombinationen werden Merkmale ermittelt, die diese Dokumente voneinander unterscheiden. Abschließend bewerten die Teilnehmer alle Dokumente anhand der zuvor identifizierten Konstrukte. In einem zweiten Schritt erhält eine andere Gruppe von Teilnehmern die Aufgabe inhaltlich ähnliche Konstrukte zusammenzufassen. Es zeigt sich, dass individuelle Konstrukte und Konstruktsysteme das Relevanzverhalten der Nutzer nur bedingt erklären können, was Howard (1994, S. 180) sowohl auf persönliche Merkmale und Motive als auch auf kontextuelle Einflüsse zurückführt.

Eine der umfangreichsten Studien zum Thema Relevanzwahrnehmung befasst sich schwerpunktmäßig mit dem Einfluss individueller Unterschiede auf die Relevanzbeurteilung (Saracevic et al., 1988; Saracevic u. Kantor, 1988a; Saracevic u. Kantor, 1988b). Hierzu werden 200 von Suchspezialisten durchgeführte Suchen zu 40 verschiedenen individuell geäußerten Informationsbedürfnissen auf Überschneidungen beim Suchen und Auffinden relevanter Dokumente durch unterschiedliche Suchende analysiert. Insgesamt zeigen sich innerhalb der Suchen zu denselben Informationsbedürfnissen nur geringe Überschneidungen, sowohl was die verwendeten Suchbegriffe, als auch die gefundenen Dokumente betrifft. So beträgt die Überschneidung relevanter Dokumente bei fünf Suchenden nur 18% (Saracevic u. Kantor, 1988b, S. 204). Für die Messung von Relevanz bedeutet dies vor allem zweierlei. Einerseits wird einmal mehr deutlich, wie wichtig es ist, die Nutzer in die Evaluierung der Qualität von Suchsystemen einzubeziehen und über Batch-Evaluierungen hinaus zu denken. Zum anderen zeigen die Untersuchungsergebnisse erneut, wie komplex und vielschichtig das Konstrukt der Relevanz ist, sodass eine isolierte Betrachtung der einzelnen Einflüsse kaum durchführbar erscheint.

Wang und Soergel (1998) und Wang und White (1999) führen eine Längsschnittstudie mit dem Ziel durch, individuelle Relevanzkriterien unter realen Bedingungen zu untersuchen. Im ersten Teil der Studie liegt der Fokus auf Entscheidungen im Zusammenhang mit der Selektion relevanter Dokumente (Wang u. Soergel, 1998). Die Inhaltsanalyse der Think-Aloud-Protokolle ergibt, dass die Themenzugehörigkeit von allen genannten Relevanzkriterien mit 65% der Kommentare am häufigsten als Grund für die Auswahlentscheidung angeführt wird. Darüber hinaus können sechs Entscheidungsregeln identifiziert werden, die die Probanden anwenden, um zu beurteilen, ob ein Dokument relevant ist oder nicht (ebd., S. 127 f.). Im zweiten Teil der Studie, in dem es primär um Entscheidungen geht, die sich auf das Lese- und Zitierverhalten der Probanden

beziehen, können Wang und White (1999, S. 106 ff.) die generelle Gültigkeit dieser Entscheidungsregeln erneut bestätigen. Im Folgenden werden die betreffenden Entscheidungsregeln kurz erläutert:

Ausscheidungsregel (elimination rule) – Diese Regel findet in der Erhebung von Wang und Soergel (1998) am häufigsten (in 382 von 778 betrachteten Fällen) Anwendung. Nutzer, die diese Regel einsetzen, gehen davon aus, dass ein einziges Kriterium ausreicht, um das gesamte Dokument negativ beurteilen zu können. Laut Wang und Soergel (ebd.) spiegelt diese Regel das Prinzip des geringsten Aufwandes wider, da die Informationsverarbeitung abgebrochen werden kann, sobald das erste unerwünschte Merkmal identifiziert wird.

Multikriterielle Regel (multiple-criteria rule) – Auch diese Entscheidungsregel kommt im Rahmen der Untersuchung von (ebd.) häufig (in 323 Fällen) zum Einsatz. Im Unterschied zu der ersten Regel wägen Nutzer dieser Regel zunächst mehrere Kriterien ab, bevor sie sich für Akzeptanz oder Ablehnung eines Dokuments entscheiden. Diese Nutzer möchten also mit einer gewissen Sicherheit sagen können, dass die getroffene Entscheidung richtig ist.

Kettenregel (chain rule) – Mit deutlichem Abstand, aber mit der dritthäufigsten Nennung (etwa die Hälfte der Nutzer wenden diese Regel in 49 der betrachteten Fälle an) wird die Kettenregel verwendet. Diese Auswahlregel besagt, dass Aufsätze, die in einer logischen Beziehung zueinander stehen, relevant sind. Beispiele für diese Regel sind Zitationen, Kritiken sowie Beiträge in einem Sammelband oder Sonderheft einer Zeitschrift zu einem einzelnen Thema (ebd., S. 128).

Satisficing Regel (satisfice rule) – Satisficing setzt sich aus den englischen Begriffen *satisfying* (=zufriedenstellend) und *suffice* (=genügen) zusammen und bezeichnet in der Entscheidungstheorie die Strategie, nicht nach der bestmöglichen, sondern nach einer genügend guten Lösung zu suchen. Diese Auswahlregel wenden vier Nutzer in neun Fällen an, wenn sie das Gefühl haben eine ausreichende Menge an Dokumenten gefunden zu haben (ebd.).

Dominanzregel (dominance rule) – Eine weitere Entscheidungsregel, die von vier Nutzern in acht Fällen angewendet wird, ist die Dominanzregel. Nach dieser Regel wird ein Dokument abgelehnt, sobald es ein anderes gibt, das in allen Merkmalen gleichwertig ist, jedoch bei mindestens einem Merkmal einen Vorteil aufweist. Die Verwendung dieser Regel macht deutlich, dass Nutzer einen effizienten Arbeitsprozess wünschen, im Rahmen dessen nicht alle relevanten Dokumente notwendigerweise bearbeitet werden müssen.

Angebotsknappheitsregel (scarcity rule) – Die Regel der Angebotsknappheit wird ebenfalls in vergleichsweise wenigen Fällen genutzt. Wang und Soergel (ebd.) stellen fest, dass einige Nutzer flexibel auf die jeweilige Ergebnismenge reagieren und ihre Relevanzkriterien dementsprechend anpassen. So lockern vier der Probanden in sieben Fällen ihre Auswahlregeln im Fall nur weniger zurückgelieferter Dokumente, um zumindest einige Dokumente als relevant abspeichern zu können.

Die hier vorgestellten Entscheidungsregeln liefern eine Handlungssystematik, die es erlaubt, das Auswahlverhalten der Nutzer zu klassifizieren. Interessant ist in diesem Kontext insbesondere

die Frage, wie in der Dynamik des Suchprozesses unterschiedliche Regeln angewendet werden bzw. ein Regelwechsel stattfindet. Diese Prozesse und weitere Verhaltensweisen, die von der Güte der präsentierten Ergebnisse abhängen, werden im folgenden Abschnitt besprochen.

3.1.3. Dynamische Aspekte bei der Relevanzbeurteilung

Die dieser Arbeit zugrunde gelegte situative Relevanzdefinition impliziert bereits, dass die oben genannten Regeln und Kriterien einer zeitlichen Dynamik unterliegen und eine Reihe von äußeren Faktoren die Meinungsbildung sowie das endgültige Relevanzurteil beeinflussen. Darauf Bezug nehmend erklärt Harter (1992, S. 612): „ [. . .] relevance judgments are a function of one's mental state at the time a reference is read. They are not fixed; they are dynamic. " Im Folgenden werden ausgewählte Studien vorgestellt, die diese Dynamik und den situativen Kontext genauer analysieren.

Scholer und Turpin (2008) und Scholer et al. (2008) gehen von der Existenz individueller Relevanzschwellenwerte aus, die von Person zu Person variieren und es so problematisch erscheinen lassen, system- und benutzerorientierte Untersuchungsergebnissse vergleichbar zu machen. Um diese Vermutung zu überprüfen, führen Scholer und Turpin (2008) zunächst eine Studie durch, in der sie das Relevanzverhalten von 11 Testpersonen vergleichen. Dazu erzeugen sie künstliche Ergebnislisten, die sich in der Relevanz des ersten dargestellten Dokuments in Bezug auf eine 4-stufige Relevanzskala unterscheiden, was einer Variation der $P@1$ entspricht. Die verwendeten Abstufungen entsprechen den Kategorien: *completely relevant*, *highly relevant*, *marginally relevant* und *not relevant*. Als abhängige Variable wird die Zeit protokolliert, die von den Testpersonen benötigt wird, um das erste Dokument als relevant zu speichern. Die Idee besteht darin, dass Probanden mit einem sehr hohen Relevanzschwellenwert nur Dokumente der höchsten Relevanzkategorie akzeptierten und damit bei Ergebnislisten mit geringerer $P@1$ länger benötigen, bis sie das erste von ihnen als relevant wahrgenommene Dokument gespeichert haben. Die Ergebnisse deuten zwar daraufhin, dass die Probanden tatsächlich unterschiedliche Relevanzschwellen besitzen, allerdings ist die betrachtete Stichprobe mit 11 Testpersonen zu klein, als dass die Erkenntnisse als repräsentativ gelten könnten. In einer weiteren Studie versuchen Scholer et al. (2008) die individuellen Relevanzschwellenwerte der Testteilnehmer experimentell zu bestimmen. Dazu bewerten 40 Testpersonen insgesamt 60 Dokumente zu drei TREC-Themen auf einer binären Relevanzskala. Die korrespondierenden Jurorenurteile liegen auf einer 3-stufigen Skala von *nicht relevant* über *relevant* bis *sehr relevant* vor. Mit dem Ziel die Nutzer in unterschiedliche Kategorien bezüglich ihres Relevanzschwellenwertes einzuteilen, erproben Scholer et al. (ebd.) verschiedene Ansätze, um Juror- und Nutzerurteile zu vergleichen. Insgesamt zeigen die Ergebnisse, dass sich die Relevanzkriterien der Benutzer stark unterscheiden: „Our results indicate that relevance thresholds vary significantly between individuals; that is, some searchers have a low tolerance for documents that are only of marginal use to an information need, while others consider these to be useful to the same search request." (ebd., S. 48) Dies lässt die Autoren vermuten, dass individuelle Abweichungen in Bezug auf Relevanzschwellenwerte eine Ursache für die Unterschiede zwischen system- und benutzerorientierten Untersuchungen darstellen könnten (vgl. Abschn. 3.2.1). Auch Wang und Soergel (1998) stellen fest, dass die Schwelle, ab wann ein Suchergebnis als relevant eingestuft wird, stark personenabhängig sein kann (vgl. Abschn. 3.1.2). Die Tatsache, dass einige Teilnehmer ihre Relevanzkriterien lockern, wenn nur wenige

relevante Ergebnisse zur Verfügung stehen, lässt Wang und Soergel (1998, S. 128) vermuten, dass der jeweilige situative Kontext eine wichtige Rolle bei der Festlegung der individuellen Relevanzschwelle des einzelnen Nutzers spielt. Wang und White (1999, S. 101 f.) berichten im zweiten Teil ihrer Studie von einem Fall, in dem einer der Teilnehmer für die Auswahl- und Leseentscheidung eine niedrigere Relevanzschwelle ansetzt als für die anschließende Zitierentscheidung. In diesem Zusammenhang wird eine weitere Kontextbedingung sichtbar, die zu einem Anheben oder Senken der individuellen Relevanzschwelle beitragen kann. Neben der aktuellen Verfügbarkeit relevanter Suchergebnisse kann auch die jeweilige Phase im Evaluierungsprozess Grund für eine Anpassung relevanzbezogener Maßstäbe sein. Kuhlthau (1993a) diskutiert das Verhalten von Nutzern in den verschiedenen Phasen des Suchprozesses und führt derartige Verhaltensänderungen auf einen vorübergehenden Zustand der Unsicherheit zurück, der sich mit zunehmendem Wissen verringert.

Ähnliche Anpassungsreaktionen werden auch von Smucker und Jethani (2010a) beschrieben. Die Autoren untersuchen den Zusammenhang zwischen System- und Benutzerleistung in einem Within-Subject-Design (vgl. Abschn. 3.2.1). Die 48 Testpersonen bearbeiten acht Aufgaben aus dem TREC 2005 Robust Track, je vier pro Testphase. Während die Teilnehmer in der ersten Phase gebeten werden, jedes Dokument in der Ergebnisliste zu bewerten, können sie in der zweiten Phase frei entscheiden, welche Dokumente sie ansehen und bewerten wollen. Die Ergebnislisten werden so manipuliert, dass die Precision für das gute System $0{,}6$ beträgt, für das schlechte System hingegen $0{,}3$. Auch hier zeigt sich ein adaptives Verhalten der Teilnehmer. Nutzer des besseren Systems neigen in beiden Testphasen dazu, strengere Relevanzkriterien auf bessere Ergebnislisten anzuwenden.

Die Mehrzahl der Studien, die sich mit dem Relevanzverhalten von Nutzern beschäftigen, sind prozessorientiert und als Längsschnitt mit mehreren Messzeitpunkten angelegt. Bruce (1994) beobachtet sechs Studenten bei Recherchearbeiten mit dem Ondisc-Informationssystem. In der Studie werden Entwicklungen im Bewertungsverhalten der Testteilnehmer über drei Zeitpunkte (vor, während und nach der Interaktion) hinweg erhoben. Dazu bewerten die Teilnehmer die relative Wichtigkeit verschiedener Dokumenteigenschaften im Hinblick auf ihr Relevanzurteil mittels der *Magnitude-Estimation-Methode* (vgl. Abschn. 4.2.2.1). Obwohl die untersuchte Stichprobe nicht repräsentativ sein dürfte, kann Bruce (ebd.) zeigen, dass die vorgeschlagene Methode prinzipiell in der Lage ist, dynamische Aspekte im Bewertungsverhalten einzelner Nutzer messbar und sichtbar zu machen. Auch weitere Studien befassen sich mit der Beobachtung von zeitlichen Entwicklungen im Bewertungsverhalten (Tang u. Solomon, 1998; Tang u. Solomon, 2001; Smithson, 1994; Vakkari u. Hakala, 2000; Vakkari, 2001). Zu diesem Zweck begleiten Tang und Solomon (1998) eine Studentin während ihrer Recherche für eine Hausarbeit. In zwei zeitlich auseinanderliegenden Interviewphasen soll die Probandin die gefundenen Dokumente bewerten und relevante Passagen markieren. Es kann ein Lerneffekt im Sinne eines veränderten Bewertungsverhaltens sowie ein gesteigertes Selbstvertrauen in die eigenen Relevanzurteile von der ersten auf die zweite Interviewphase festgestellt werden (ebd., S. 245 ff.). In zwei weiteren von Tang und Solomon (2001) durchgeführten Untersuchungen ergeben sich ähnliche Ergebnisse. In beiden Studien werden Veränderungen des Bewertungsverhaltens der Probanden zu zwei Messzeitpunkten (bibliographische Referenzen und Volltextdokumente) miteinander

verglichen. Allerdings verwenden die hier genannten Studien sehr unterschiedliche methodische Ansätze, sodass ein direkter Vergleich der Ergebnisse nicht möglich ist (ebd., S. 680 ff.). In der ersten Studie, die als Laborexperiment durchgeführt wird, evaluieren 90 Teilnehmer 20 Dokumente unter Verwendung 15 vordefinierter Relevanzkriterien. In der zweiten Studie, die als Feldbeobachtung umgesetzt ist, suchen neun Teilnehmer nach Dokumenten, die ihre eigenen Forschungen voranbringen. Gleichwohl belegen beide Studien erneut, dass Relevanzkriterien einer Dynamik unterliegen und sich im Zeitverlauf sukzessive ändern. Die von Smithson (1994) skizzierte Längsschnittstudie bezieht neben den Relevanzbewertungen der Testpersonen auch deren Zufriedenheitsurteile in die Analyse ein. Die erzielten Ergebnisse variieren sowohl zwischen den einzelnen Probanden als auch mit der Zeit zum Teil erheblich. Darüber hinaus kann Smithson (ebd., S. 218 f.) eine deutliche Abhängigkeit der Ergebnisse von den gewählten Evaluierungsmaßen nachweisen. Vakkari und Hakala (2000) und Vakkari (2001) führen eine Längsschnittstudie zur Konkretisierung von Kuhlthaus Modell des Informationssuchprozesses (Kuhlthau, 1993b) für die Domäne des Information Retrieval durch. Dieses Modell geht davon aus, dass das Such- und Nutzungsverhalten entscheidend davon abhängt, in welcher Prozessphase sich ein Nutzer gerade befindet. Die hier vorgestellte Untersuchung von Vakkari und Hakala (2000) und Vakkari (2001) analysiert zeitliche Entwicklungen im Suchverhalten zu drei Messzeitpunkten und legt einen systematischen Zusammenhang zwischen der jeweiligen Handlungsphase im Suchprozess und den darin verwendeten Suchanfragen und -strategien nahe (ebd., S. 301 ff.). Im Gegensatz zu den anderen Studien ergibt sich jedoch keine wesentliche Veränderung in Bezug auf die verwendeten Relevanzkriterien (Vakkari u. Hakala, 2000, S. 551).

Neben diesen relevanzspezifischen Anpassungseffekten können bei der Relevanzbewertung auch psychologische Effekte wie *Antworttendenzen* (vgl. Abschn. 4.2.3.2) auftreten. Als Antworttendenzen werden personenspezifische Antwortmuster bezeichnet, die bei Befragungen auftreten können (Bortz u. Döring, 2006; Jonkisz et al., 2008). Eine der bekanntesten Antworttendenzen ist die *Akquieszenz* (auch Zustimmungstendenz), also die Neigung Fragen unabhängig von ihrem Inhalt eher zustimmend zu beantworten (Bortz u. Döring, 2006, S. 236). Weitere Antworttendenzen, die bei der Relevanzbewertung leicht auftreten und die Aussagekraft der Ergebnisse unter Umständen beeinträchtigen können, sind die zur Akquieszenz gegenläufige *Ablehnungstendenz* sowie die *Tendenz zur Mitte* (ebd., S. 236). Allerdings können diese Effekte durch eine sorgfältige Fragebogenkonstruktion im Vorfeld der Studien verringert bzw. vermieden werden (Jonkisz et al., 2008; Raab-Steiner u. Benesch, 2010). Eine mögliche Strategie zur Vermeidung derartiger Verhaltensmuster liegt in der Verwendung graduell abgestufter Antwortmöglichkeiten, die dichotome Fragen vermeiden. Hinsichtlich der Relevanzbewertung in kontrollierten IR-Experimenten stellt sich somit die ganz wesentliche Frage, inwiefern binäre Relevanzurteile von diesen Verhaltenstendenzen betroffen sind und wie gegebenenfalls Verfälschungen der Messergebnisse durch Antworttendenzen vermieden oder vermindert werden können. Diese Fragestellung wird in Abschnitt 4.2 im Rahmen der Operationalisierung der untersuchten Variablen und der Kontrolle von Störvariablen wieder aufgegriffen und ausführlich diskutiert.

Über diese mit dem Testdesign verbundenen Aspekte hinaus, zeigt dieser Abschnitt, wie sich die im Rahmen der theoretischen Relevanzkonstrukte postulierten dynamischen und kontextuellen Abhängigkeiten in interaktiven Nutzerstudien zum Informationssuchverhalten manifestieren und

nachgewiesen werden können. In Bezug auf das hier angestrebte Forschungsvorhaben erscheinen insbesondere die beschriebenen Anpassungseffekte aufgrund unterschiedlicher Systemgüten relevant. Der nun folgende Abschnitt hingegen beschäftigt sich mit der bereits mehrfach angesprochenen Übertragbarkeit systemorientierter IR-Studien auf den Nutzerkontext sowie der Frage nach der Wahrnehmung des eigenen Sucherfolgs.

3.2. Die Bewertung des individuellen Sucherfolgs

Wie in der Einleitung dargestellt existieren hinsichtlich der Evaluierung von Suchergebnissen grundsätzlich zwei unterschiedliche Bewertungsansätze: einerseits die in der Tradition der Cranfield-Experimente stehenden systemorientierten Verfahren und anderseits die stärker am Benutzer ausgerichteten Verfahren (vgl. Kap. 1). Die bisherigen Ausführungen in Kapitel 2 haben gezeigt, dass benutzerorientierte Verfahren von einer Vielzahl subjektiver Einflussfaktoren wie z.B. der Art der Informationsverarbeitung, den verschiedenen Vorerfahrungen der Nutzer sowie motivationalen und situativen Komponenten beeinflusst werden. Mit Blick auf die im vorherigen Abschnitt vorgestellten Forschungsergebnisse in Bezug auf Verhaltensänderungen bei der Relevanzbeurteilung (vgl. Abschn. 3.1.3) ist im Rahmen der Beurteilung des individuellen Sucherfolgs davon auszugehen, dass besonders jene Aspekte des Suchprozesses eine Rolle spielen, die den Bedürfnissen und Erwartungen des Benutzers entsprechen. Die dadurch erhöhte Komplexität des Evaluierungsverfahrens hat in der IR-Forschung dazu geführt, dass die Evaluierung von IR-Systemen in natürlichen Nutzungskontexten lange Zeit vernachlässigt wurde. Da jedoch der Sucherfolg, den reale Benutzer mit IR-Systemen erzielen, am Ende über deren Anwendbarkeit entscheidet, ist die Weiterentwicklung benutzerorientierter Verfahren für die IR-Forschung von zentraler Bedeutung. Järvelin und Ingwersen (2004) fassen diese Situation wie folgt zusammen: „The real issue in information retrieval systems design is not whether its recall-precision performance goes up by a statistically significant percentage. Rather, it is whether it helps the actor solve the search task more effectively or efficiently.“

Vor diesem Hintergrund ergeben sich zwei wichtige Fragen. Die erste betrifft die Problematik, ob bessere Suchsysteme den objektiv messbaren Sucherfolg des Benutzers erhöhen und, wenn ja, ob sich die ermittelten Leistungswerte mit den im Kontext systemorientierter Verfahren festgestellten Unterschieden erklären lassen. Die zweite Frage, besteht darin, ob bessere Suchsysteme den subjektiv wahrgenommenen Nutzen der Suchergebnisse für die Befriedigung eines individuellen Informationsbedürfnisses erhöhen. Diese Erfolgsdimension ist eng mit der in Abschnitt 3.3 besprochenen Zufriedenheit der Benutzer verknüpft und diesbezügliche Wahrnehmungen gehen in der Realität häufig ineinander über. Im weiteren Verlauf dieses Kapitels werden dennoch beide Wahrnehmungsdimensionen zunächst getrennt voneinander betrachtet, um die spezifischen Eigenheiten und die Funktionsweise dieser beiden Leistungsbewertungen besser zu verstehen. Aus diesem Grund werden in Abschnitt 3.2.1 vorerst Forschungsergebnisse dargestellt, die die Übertragbarkeit systemorientierter Evaluierungsergebnisse auf den Benutzerkontext thematisieren. Bevor in Abschnitt 3.3 mit der Zufriedenheit eine den gesamten Suchprozess in den Blick nehmende Sicht auf die Suchergebnisse eingenommen wird, stellt Abschnitt 3.2.2 noch einige Studien vor, die den wahrgenommenen Sucherfolg primär auf Ebene der zurückgelieferten Suchergebnisse erfassen.

3.2.1. Übertragbarkeit systemorientierter Evaluierungsergebnisse

Trotz der zentralen Bedeutung der Benutzerperspektive und der Benutzerzufriedenheit dominieren leistungsorientierte Sichtweisen die wissenschaftliche Diskussion um die Evaluierung von Retrievalergebnissen. Zu nennen sind in diesem Zusammenhang insbesondere die Studien von Hersh et al. (2000), Turpin und Hersh (2001), Allan et al. (2005) und Turpin und Scholer (2006), die sich primär mit der Frage befassen, inwieweit sich Ergebnisse aus systemorientierten Evaluierungen auf reale Benutzer und deren individuelle Informationsbedürfnisse übertragen lassen. Genauer untersuchen diese Autoren, ob eine Korrelation zwischen System- und Benutzerleistung nachgewiesen werden kann. Grundsätzlich wird in diesem Zusammenhang die Benutzerleistung mit ähnlichen Methoden wie in der systemorientierten Evaluierung erfasst (vgl. Abschn. 4.2.2.2). Die zugrunde liegende Fragestellung lautet jedoch diesmal: Wie gut sind verschiedene Benutzer in der Lage, mit einem zu evaluierenden System relevante Dokumente zu finden? Um die Übertragbarkeit systemorientierter Ergebnisse auf den Benutzerkontext zu untersuchen, wird wie im systemorientierten Fall eine vollständig bewertete Testkollektion benötigt. Deren Erstellung ist im benutzerorientierten Fall jedoch aufwändiger und erfordert u.a., dass die jeweiligen Informationsbedürfnisse im Sinne der in Abschnitt 2.4 erwähnten *simulierten Arbeitsaufgaben* in einen narrativen Kontext eingebettet werden (vgl. Abschn. 4.1.3.2). Darüber hinaus muss interaktionsinduzierten Veränderungen der Relevanzbeurteilung Rechnung getragen werden (vgl. Abschn. 3.1.3), was z.B. durch graduell abgestufte Relevanzurteile erreicht werden kann (vgl. Abschn. 4.2.1.1). Zwei wesentliche Erkenntnisse sind für den vorliegenden Zusammenhang von Bedeutung: Zum einen, dass Benutzer insbesondere in Bezug auf recallorientierte Leistungsmaße häufig in der Lage zu sein scheinen, systemseitige Leistungsunterschiede zu kompensieren. Zum anderen, dass der Prozess der Relevanzbeurteilung oft kontextabhängig ist und seine Generalisierbarkeit weiter erforscht werden muss (vgl. Abschn. 3.1.3). Im Folgenden werden die Ergebnisse ausgewählter Studien vorgestellt und diskutiert.

Während einige Studien einen statistisch signifikanten Einfluss der Suchergebnisqualität auf die Benutzerleistung identifizieren können (Allan et al., 2005; Al-Maskari et al., 2008b; Smucker u. Jethani, 2010a), kommen andere Untersuchungen zu dem Schluss, dass Benutzer zumindest partiell dazu in der Lage sind, derartige Qualitätsunterschiede auszugleichen (Turpin u. Hersh, 2001; Turpin u. Scholer, 2006; Smith u. Kantor, 2008). Jedoch sind in allen Fällen der genaue Untersuchungsaufbau und die tatsächlich verwendeten Systemqualitätsunterschiede, unter welchen die jeweiligen Ergebnisse erhalten werden, zu berücksichtigen.

Die erste Studie, die in diesem Rahmen vorgestellt wird, stammt von Allan et al. (2005). Die Autoren verwenden die *Binary Preference*[1] (BPref) um in ihrer Untersuchung die Qualität der Suchergebnislisten zu manipulieren (vgl. Abschn. 4.2.1.1). Konkret werden in dem Experiment Listen in acht verschiedenen BPref-Abstufungen zu 45 Suchthemen generiert und es werden 33 Probanden rekrutiert. Die zentrale Aufgabe der Teilnehmer besteht darin, aus diesen vordefinierten Suchergebnislisten möglichst viele unterschiedliche relevante Antwortfacetten zu extrahieren. Da Allan et al. (ebd.) zur Bewertung der Benutzerleistung neben dem sog. *instance recall*, d.h. dem Anteil gefundener korrekter Antwortfacetten, auch die zur Aufgabenerledigung benötigte Zeit betrachten, wird keine zeitliche Begrenzung für die Bearbeitung der einzelnen Aufgaben

[1] Bei der Binary Preference handelt es sich um ein Systemleistungsmaß, das auswertet, wie oft im Mittel relevante vor irrelevanten Dokumenten zurückgegeben werden (vgl. Abschn. 4.2.1.1).

vorgegeben. Im Ergebnis zeigen Allan et al. (2005), dass eine erhöhte Suchergebnisqualität die Benutzerleistung signifikant beeinflussen kann. So benötigen die Teilnehmer insgesamt weniger Zeit, um die gestellten Aufgaben zu lösen und erreichen teilweise auch höhere Instance-Recall-Werte. Genauer kann ein signifikanter Einfluss der Systemleistung ab einem relativen Systemunterschied[2] von 60% in Bezug auf BPref nachgewiesen werden. Für einen relativen Unterschied von bis zu 40% hingegen ist kein Einfluss der Systemleistung feststellbar (BPref 50 vs. 70). Im Rahmen der Experimente 2 und 3 dieser Arbeit bewegt sich der relative Unterschied in Bezug auf BPref hingegen im Bereich von 50% (vgl. Abschn. 6.3.1).

Einen signifikanten Einfluss der Systemleistung auf die Benutzerleistung können auch Al-Maskari et al. (2008b) bzw. Al-Maskari et al. (2008a) nachweisen. Die 56 Teilnehmer bearbeiten dabei acht Suchaufgaben, wobei sie frei mit dem Suchsystem interagieren können. Ohne ihre Kenntnis verwenden sie dabei für vier Aufgaben ein besseres und für vier Aufgaben ein schlechteres Suchsystem. Das Zeitlimit liegt bei sieben Minuten pro Aufgabe. Die Kontrolle der Systemleistung beruht dabei auf einem Pretest, bei dem für jedes Topic drei Suchmaschinen verglichen werden. Abhängig von ihrem Abschneiden wird pro Topic jeweils die beste und schlechteste Suchmaschine für die entsprechende Aufgabe gewählt. Im Durchschnitt erreicht das schlechtere System eine *Average Precision*[3] (AvP) von $0{,}05$ das bessere System hingegen eine AvP von $0{,}20$ (vgl. Abschn. 4.2.1.1). Allerdings variiert dieser Unterschied stark in Bezug auf die einzelnen Aufgaben. Weiterhin ist kritisch anzumerken, dass die tatsächlich im Experiment auftretenden AvP-Werte nicht noch einmal überprüft werden. In diesem Sinne ist also eine präzise Kontrolle der Systemunterschiede nicht gegeben. Auch Al-Maskari et al. (2008b) finden einen signifikanten Einfluss der Systemleistung in Bezug auf die Zeit zum Auffinden des ersten relevanten Dokuments und den Benutzerrecall, der als Anzahl der gefundenen relevanten Dokumente operationalisiert ist. Weiterhin beziehen sie für den Recall die Rankingposition der Dokumente ein und weisen nach, dass der Unterschied im Wesentlichen innerhalb der ersten zehn Rankingpositionen zu beobachten ist. Der mittlere absolute Systemunterschied bewegt sich mit $0{,}15$ auf einem ähnlichen Niveau wie bei den im Rahmen dieser Arbeit durchgeführten Nutzerexperimenten (vgl. Abschn. 5.3.1, 6.3.1 u. 7.3.1). Allerdings spielt sich der Vergleich am unteren Ende der Systemleistungsskala ($0{,}05$ vs. $0{,}20$) ab, was zu weitaus größeren relativen Systemunterschieden führt. Ein signifikanter Einfluss der Systemgüte kann für mittlere relative Systemunterschiede[2] von mindestens 200% (AvP $0{,}05$ vs. AvP $0{,}15$) nachgewiesen werden, was für solch einen hohen relativen Unterschied auch plausibel erscheint. Für Topics mit einem relativen Systemunterschied von ca. 30% (AvP $0{,}11$ vs. AvP $0{,}14$) hingegen finden Al-Maskari et al. (ebd.) keinen signifikanten Einfluss der Systemleistung auf den Nutzerrecall, was in etwa dem relativen Systemunterschied von 35% entspricht, der in dieser Arbeit betrachtet wird (vgl. Abschn. 6.3.1 u. 7.3.1).

Auch Turpin und Hersh (2001) gehen in zwei Studien der Frage nach, ob system- und benutzerorientierte Evaluierungen zu denselben Ergebnissen führen. Die erste Studie wird im Rahmen des

[2] Der relative Systemunterschied zwischen zwei Suchsystemen S_1 und S_2 in Bezug auf ein Systemleistungsmaß L ist als $\frac{L(S_2)-L(S_1)}{L(S_1)}$ definiert, wobei diese Zahl mit 100 zu multiplizieren ist, um die berichteten Prozentzahlen zu erhalten (Sanderson u. Zobel, 2005). Damit diese Größe positiv ist, wird die Konvention getroffen, dass $L(S_1)$ größer als $L(S_2)$ ist. Damit beschreibt der relative Systemunterschied die Verbesserung der Systemleistung des besseren im Vergleich zum schlechteren System.

[3] Bei der Average Precision handelt es sich um ein Systemleistungsmaß, das sich als Mittelwert über die Precision der Ergebnisliste ausgewertet an allen Rankingpositionen mit relevanten Dokumenten ergibt (vgl. Abschn. 4.2.1.1).

TREC-8 interactive track durchgeführt und besteht wie die Studie von Allan et al. (2005) in einer Instance-Recall-Aufgabe. Die zweite Studie hingegen beinhaltet Question-Answering-Aufgaben des TREC-9 interactive track, bei welchen die Testpersonen die korrekten Antworten auf natürlichsprachliche Fragen auffinden sollen. Für beide Experimente verwenden die Autoren zwei Suchsysteme mit unterschiedlicher *Mean Average Precision*[4] (MAP) (TREC-8: 0.27 vs. 0.32 MAP; TREC-9: 0.27 vs. 0.35 MAP, vgl. Abschn. 4.2.1.1). Im Vergleich zu den zuvor diskutierten Studien fällt der relative Systemunterschied mit knapp 20% hier also weit geringer aus. Wie in der zuvor beschriebenen Studie von Al-Maskari et al. (2008b) wird auch in diesem Fall die Systemleistung in einem Pretest ermittelt und während der Interaktion der Probanden mit dem Testsystem nicht kontrolliert. Allerdings überprüfen Turpin und Hersh (2001) a posteriori, dass die an die Probanden ausgelieferten Rankinglisten tatsächlich den gewünschten Qualitätsunterschied aufweisen, wobei diese interessanterweise mit einem relativen Unterschied von 47% und 68% im Mittel noch größer ausfallen als im Pretest. Während die Bearbeitungszeit für die Instance-Recall-Aufgaben 20 Minuten beträgt, haben die Teilnehmer in dem Question-Answering-Experiment jeweils fünf Minuten Zeit, die Aufgaben zu lösen. Wie aufgrund des geringen relativen Systemunterschiedes zu vermuten, kann in keiner der beiden Studien ein signifikanter Einfluss der Systemqualität auf die Benutzerleistung beobachtet werden. Weiterhin zeigen Turpin und Hersh (ebd.), dass Benutzer die systembedingten Leistungsunterschiede dadurch ausgleichen, dass sie bspw. mehr Suchanfragen stellen und mehr Dokumente aufrufen. Die Kompensation des Leistungsunterschiedes geht also auf Kosten der Effizienz.

Nachdem die beiden zuletzt vorgestellten Studien Beispiele für Untersuchungen sind, in welchen zugunsten einer natürlicheren Nutzerinteraktion eine geringere Kontrolle der Systemqualität gewählt wird, soll im Folgenden mit der Studie von Turpin und Scholer (2006) noch ein Untersuchungsdesign beschrieben werden, das versucht, beiden Anforderungen gerecht zu werden. Die Probanden können zwar frei mit dem Suchsystem interagieren und eigene Suchanfragen formulieren, die zurückgelieferten Ergebnislisten werden jedoch aus einem Pool vorgenerierter Ergebnislisten mit fester Systemleistung ausgewählt. Auf dieses Vorgehen wird auch für die Benutzertests im Rahmen der vorliegenden Arbeit zurückgegriffen (vgl. Abschn. 4.1.3.1 u. 4.2.1.1). Als Testkollektion dienen Web-Track-Daten der Evaluierungsinitiative TREC. Turpin und Scholer (ebd.) legen den Nutzertest als Messwiederholungstudie an, bei dem die 30 Probanden für jede der fünf betrachteten Systemgüten jeweils 10 Aufgaben bearbeiten. Dabei wird die Systemgüte anhand der MAP erfasst, die über die Abstufungen 0,55, 0,65, 0,75, 0,85 und 0,95 variiert. Die Ausgangsthese von Turpin und Scholer (ebd.) ist, dass der Schwierigkeitsgrad der Suchaufgaben in vorherigen Studien eine mögliche Erklärung für fehlende Korrelationen zwischen System- und Benutzerleistung gewesen sein könnte. Infolgedessen legen sie für ihre eigene Studie besonderen Wert auf die Einfachheit der Suchaufgaben. Die Teilnehmer sollen deshalb innerhalb von fünf Minuten so viele relevante Antwortdokumente wie möglich zu dem vorgegebenen Suchthema finden. Im Ergebnis kann kein signifikanter Einfluss der Systemleistung für die Zeit zum Auffinden des ersten relevanten Dokuments nachgewiesen werden. Für den Recall, gemessen als die Anzahl der gefundenen relevanten Dokumente, ergibt sich ein signifikanter Unterschied zwischen den Systemgüten 0,55 und 0,75 sowie 0,65 und 0,75. Der Effekt fällt mit im Mittel 0,3 Dokumenten

[4] Bei der Mean Average Precision handelt es sich um ein Systemleistungsmaß, das sich als Mittelwert der AvP über unterschiedliche Suchanfragen ergibt (vgl. Abschn. 4.2.1.1).

jedoch relativ gering aus. Insgesamt kann also nur ein geringer Einfluss der unterschiedlichen MAP-Abstufungen auf die Benutzerleistung festgestellt werden.

Während die bis hierher berichteten Studien primär der Frage nachgehen, ob eine Korrelation zwischen System- und Benutzerleistung nachgewiesen werden kann, sei zuletzt noch auf eine Gruppe von Studien hingewiesen, die mögliche Gründe für einen geringen bzw. nicht vorhandenen Benutzerleistungszuwachs untersuchen. Dies ist insbesondere vor dem Hintergrund interessant, dass hier auf die in den vorangegangenen Abschnitten dargelegten Konzepte zur Kontextabhängigkeit der Relevanzwahrnehmung zurückgegriffen wird. So wird z.B. in Abschnitt 3.1.3 gezeigt, dass die Relevanzbeurteilung im Benutzerkontext einer zeitlichen Dynamik unterliegt und das endgültige Relevanzurteil zudem von einer Reihe äußerer Faktoren beeinflusst werden kann. Einige der im Folgenden zitierten Studien sind in diesem Rahmen bereits in Abschnitt 3.1.3 behandelt. In diesem Abschnitt liegt der Schwerpunkt der Betrachtungen jedoch auf den Folgen dynamischer Relevanzurteile für den späteren Sucherfolg. Grob lassen sich in der Literatur drei Erklärungsansätze für die Diskrepanz zwischen system- und benutzerorientierter Evaluierungen ausmachen: Zum einen ist der Aufwand zu berücksichtigen, den ein Benutzer erbringen muss, um sein Informationsbedürfnis zu erfüllen. In diesem Zusammenhang deuten Studien darauf hin, dass Benutzer ihren Suchaufwand intuitiv zu erhöhen scheinen, wenn der gewünschte Sucherfolg zunächst ausbleibt. Zum anderen ist eine situative Anpassung der verwendeten Relevanzkriterien an die Qualität der Suchergebnisse zu beobachten. So zeigen einige Studien, dass Benutzer ihre Relevanzkriterien lockern, wenn nur eine geringe Anzahl relevanter Ergebnisse zur Verfügung steht. Darüber hinaus enthält die Relevanzwahrnehmung auch eine individuelle Komponente, die die Vergleichbarkeit system- und benutzerorientierter Evaluierungen einschränkt. Diesbezüglich ergeben Studien, dass die Schwelle, ab wann ein Suchergebnis als relevant eingestuft wird, außer von den äußeren Umständen auch vom jeweiligen Benutzer abhängig ist.

Eine erste Studie, die zeigt, wie Benutzer ihre Suchstrategien anpassen, um ihre Informationsbedürfnisse zu befriedigen, geht auf Smith und Kantor (2008) zurück (vgl. Abschn. 2.4). Die Autoren untersuchen das Suchverhalten von 36 Testpersonen in einem Untersuchungsdesign mit zwei Test- und einer Kontrollgruppe. Ihre Ausgangshypothese ist, dass Benutzer Systemunterschiede kompensieren können, weil sie ihr Verhalten situativ anpassen. Smith und Kantor (ebd.) simulieren unterschiedlich gute Retrievalsysteme, indem je nach Untersuchungsbedingung Dokumente aus den vorderen oder hinteren Rangplätzen der entsprechenden Google-Anfrage präsentiert werden. Smith und Kantor (ebd.) können ihre Ausgangshypothese bestätigen. Tatsächlich passen die Testpersonen ihre Suchstrategien an, wenn sie mit einem schlechteren System konfrontiert werden. Insbesondere fällt auf, dass die Testpersonen in diesem Fall mehr Suchanfragen stellen und prozentual eine größere Anzahl relevanter Dokumente identifizieren. Im Ergebnis sind sie somit in der Lage, wie im Rahmen der beiden von Turpin und Hersh (2001) berichteten Benutzerexperimente, genauso viele relevante Dokumente zu finden wie die übrigen Testteilnehmer. Die in Smith (2008) berichtete Analyse des Einflusses der Aufgabenschwierigkeit ergibt darüber hinaus, dass die Frequenz der gestellten Suchanfragen bei den als schwieriger eingestuften Aufgaben signifikant höher ausfällt als für die leichteren Aufgaben.

Eine weitere mögliche Anpassungsreaktion wird von Smucker und Jethani (2010a) in ihrer

Studie zum Zusammenhang zwischen System- und Benutzerleistung identifiziert (vgl. Abschn. 3.1.3). Zur Erinnerung: Die 48 Teilnehmer bearbeiten in dieser Untersuchung insgesamt acht Suchaufgaben, je vier Aufgaben mit einem besseren System und vier Aufgaben mit einem schlechteren System. Dabei beträgt die Precision der Suchergebnislisten des besseren Systems $0{,}6$ und $0{,}3$ für das schlechtere System, was einem relativen Systemunterschied von 100% entspricht. Das durch Smucker und Jethani (ebd.) beobachtete Verhalten weist auf eine situative Anpassung der verwendeten Relevanzkriterien an die vorgefundene Systemqualität hin. So scheinen Nutzer des besseren Systems zu einer strengeren Auslegung der Relevanzkriterien zu neigen, während sich Nutzer des schlechteren Systems an eine weniger strenge Auslegung zu halten scheinen. Im Gegensatz zu der Studie von Smith und Kantor (2008) können die Testpersonen in dieser Studie den Systemunterschied nicht kompensieren, was sich an den signifikant besseren Leistungen der Testpersonen mit dem besseren System zeigt, jedoch in Anbetracht des erneut recht hohen relativen Systemunterschieds durchaus plausibel erscheint. Dieses Verhalten wird auch in der zuvor beschriebenen Studie von Al-Maskari et al. (2008a) beobachtet.

Einen weiteren Aspekt stellt die Vermutung dar, dass Benutzer individuelle Schwellenwerte bezüglich der Relevanz von Dokumenten besitzen. Während manche Nutzer eher sparsam mit positiven Relevanzbewertungen umgehen, sind andere schneller bereit ein positives Relevanzurteil zu fällen. Scholer und Turpin (2009) untersuchen diese These mit einem Experiment, in welchem sie 40 Testpersonen 24 Suchaufgaben mit drei unterschiedlichen Systemen bearbeiten lassen (vgl. Abschn. 3.1.3). Die verwendete TREC-GOV2-Testkollektion beinhaltet Relevanzurteile in den Abstufungen *nicht relevant* (0), *relevant* (1) und *sehr relevant* (2). Für den Systemvergleich generieren Scholer und Turpin (ebd.) pro Suchaufgabe drei Suchergebnislisten, die sich lediglich in der Relevanz des ersten Dokumentes unterscheiden. Beim ersten System platzieren sie ein nicht relevantes Dokument auf der ersten Rankingposition, während beim zweiten ein relevantes und beim dritten ein sehr relevantes Dokument zurückgegeben wird. Die Probanden bewerten pro Aufgabe und System alle zehn Dokumente auf einer binären Relevanzskala. Entsprechend ihrer Vermutung verwenden die Autoren diese Relevanzurteile der Testpersonen im Anschluss an das Experiment, um jeden Teilnehmer einem von drei Bewertungstypen zuzuordnen, je nachdem, ob sie in der Regel Dokumente der Stufe 0, der Stufe 2 oder in Übereinstimmung mit den TREC-Juroren sowohl Dokumente der Stufe 1 als auch der Stufe 2 als relevant akzeptieren. Scholer und Turpin (ebd.) erfassen den individuellen Sucherfolg, indem sie die Zeit messen, die zum Auffinden des ersten aus Nutzersicht relevanten Dokuments nötig ist und können so zeigen, dass der durch die Systemmanipulation intendierte Qualitätsunterschied am deutlichsten im Fall der beiden zuletzt genannten Bewertungstypen zu Tage tritt.

Über die hier genannten Aspekte hinaus gibt es natürlich noch weitere mögliche Ursachen, warum die Ergebnisse system- und benutzerorientierter Evaluierungen zum Teil unterschiedlich ausfallen. Dazu gehören unter anderem unterschiedliche Vorerfahrungen, Erwartungen, Motivation, Anstrengungsbereitschaft oder Alters- und Geschlechterunterschiede, die bereits in Kapitel 2 erläutert wurden.

Zusammenfassend lässt sich also festhalten, dass Verbesserungen der Systemleistung zu einer erhöhten Nutzerleistung führen können. Allerdings hängt die Nachweisbarkeit dieses Effekts sowohl vom Untersuchungsdesign, den Testaufgaben und den betrachteten Benutzermaßen als

auch von den betrachteten Systemunterschieden und den ihnen zugrunde liegenden Systemmaßen ab. Darüber hinaus deuten die vorgestellten Studien in der Gesamtschau darauf hin, dass insbesondere der relative Systemleistungsunterschied ein guter Indikator für eine Verbesserung der Benutzerleistung darstellt. Auf der anderen Seite zeigt sich, dass geringere Systemunterschiede gerade in Bezug auf Recallmaße durch einen erhöhten Suchaufwand seitens der Nutzer kompensiert werden können. Vor diesem Hintergrund erscheint es für die in dieser Arbeit durchgeführten Nutzerstudien sinnvoll, eine große Bandbreite unterschiedlicher Benutzerleistungsmaße einzubeziehen, um unterschiedliche Aspekte des Systemeinflusses auf das Nutzerverhalten erfassen zu können.

3.2.2. Wahrnehmung des Sucherfolgs durch den Benutzer

Anhand des Vergleichs zwischen System- und Benutzerleistung in Abschnitt 3.2.1 ist deutlich geworden, dass die individuelle Wahrnehmung des Sucherfolgs durch den Benutzer eine entscheidende Rolle bei der Wahl der verwendeten Suchstrategien und der tatsächlich erreichten Suchleistung spielt. In diesem Abschnitt werden aus diesem Grund die prägnantesten Studienergebnisse zur Sucherfolgswahrnehmung vorgestellt. Studienergebnisse, welche die Wahrnehmung des gesamten Suchprozesses und ihre Bedeutung für die Gesamtzufriedenheit beschreiben, werden hingegen im darauffolgenden Abschnitt diskutiert. Im Rahmen des Übergangs von systemorientierter zu interaktiver IR-Evaluierung kann die Wahrnehmung des Sucherfolgs auf zwei Arten verstanden werden: zum einen als Fähigkeit des Benutzers, den zwischen unterschiedlichen Suchergebnislisten vorhandenen Systemunterschied wahrzunehmen und zum anderen als Fähigkeit, die selbst erbrachte Suchleistung beurteilen zu können. Studienergebnisse für beide Sichtweisen werden im Folgenden etwas näher erläutert, da sie in gewisser Weise die Grundlage für zufriedenheitsorientierte Ansätze bilden, die auf einem Soll-Ist-Vergleich zwischen erwarteter und wahrgenommener Leistung beruhen. Dabei werden zunächst zwei Studien beschrieben, die die Differenzierungsfähigkeit der Benutzer in Bezug auf unterschiedliche Systemqualitäten untersuchen und anschließend die wichtigsten auf die Wahrnehmung der eigenen Suchleistung bezogenen Studien diskutiert.

Thomas und Hawking (2006) führen vier Experimente zur Untersuchung der Wahrnehmung von Qualitätsunterschieden von Suchergebnislisten durch (vgl. Abschn. 3.3.2.2). Dazu wird in allen Experimenten ein Testsystem verwendet, das zwei Ergebnislisten nebeneinander anzeigt. Um den Einfluss der Systemleistung auf die Qualitätswahrnehmung der Probanden zu untersuchen, werden die präsentierten Ergebnislisten systematisch variiert, indem Google-Ergebnisse aus den vorderen Ranking-Plätzen Ergebnissen aus den hinteren Ranking-Plätzen gegenübergestellt werden. Die Versuchspersonen beurteilen nach jeder Suche, welche Ergebnisliste besser ist. Zusätzlich werden vier Nutzerinteraktionen erhoben: die Liste mit dem ersten aufgerufenen Dokument, die Liste mit dem letzten aufgerufenen Dokument, die Liste mit der höchsten Anzahl aufgerufener Dokumente und die Liste mit dem am höchsten gerankten aufgerufenen Dokument. Alle vier Experimente können bestätigen, dass Benutzer in der Lage sind, Qualitätsunterschiede zwischen Suchmaschinen wahrzunehmen und dementsprechend zu differenzieren. Darüber hinaus kann gezeigt werden, dass die erhobenen Interaktionen stark mit den Qualitätsurteilen der Probanden korrelieren. Zu ähnlichen Ergebnissen kommt auch die Studie von Kelly et al. (2007) (vgl. Abschn. 3.3.2.2). Die Autoren führen drei Experimente durch, in welchen sie zeigen, dass

sowohl das Ranking als auch die Precision einen signifikanten Einfluss auf die Qualitätsbeurteilung von Suchergebnislisten haben. Die Experimente sind als Within-Subject-Design umgesetzt. Dementsprechend bearbeiten alle Probanden je eine Aufgabe mit jedem der vier untersuchten Systeme, wobei die Systemleistung als unabhängige Variable fungiert. In den ersten beiden Experimenten variieren Kelly et al. (ebd.) lediglich das Ranking der aus je zehn Dokumenten bestehenden Ergebnislisten und halten die Precision konstant auf einem Wert von $0{,}5$. Im dritten Experiment verändern die Autoren hingegen die Precision, wobei sie versuchen, die Änderung der Ergebnislisten und damit des Rankings möglichst gering zu halten. Pro Suchaufgabe bewerten die Probanden die Leistung der soeben verwendeten Suchmaschine auf einer Skala von 1 (very poor) bis 10 (very good). Im Anschluss an jede durchgeführte Suche haben sie darüber hinaus die Möglichkeit, die zuvor gefällten Urteile anzupassen, um eine relative Qualitätsbeurteilung zu ermöglichen. Analog zu der zuvor besprochenen Studie legen auch die Ergebnisse dieser Untersuchung nahe, dass Benutzer Qualitätsunterschiede zwischen Suchmaschinen wahrnehmen. Außerdem wird der spezifische Einfluss von Ranking und Precision analysiert und es stellt sich heraus, dass die Precision der Ergebnislisten einen stärkeren Effekt auf die Leistungsbeurteilung ausübt als das Ranking der Dokumente. Im Kontext dieser beiden Studien kann somit gezeigt werden, dass Benutzer den Qualitätsunterschied zwischen Suchergebnislisten wahrnehmen können.

Hinsichtlich der eigenen Fähigkeit, die im Suchprozess erbrachte Leistung einzuschätzen, führen Dostert und Kelly (2009) ein interaktives IR-Experiment durch, das den Fragen nachgeht, wie und wann Benutzer die Entscheidung treffen, ihre Suche zu beenden. Als Testkollektion verwenden die Autoren Daten des TREC 2005 Robust Track. Alle 23 Probanden verwenden während des Experiments dasselbe Testsystem und bearbeiten vier Suchaufgaben, für die sie so viele relevante Dokumente wie möglich innerhalb der vorgegebenen Bearbeitungszeit von 15 Minuten identifizieren sollen. Neben der Erfassung des tatsächlichen Recalls der Probanden bitten Dostert und Kelly (ebd.) die Teilnehmer im Anschluss an das Experiment um eine persönliche Einschätzung des von ihnen erreichten Recalls. Die Ergebnisse der Studie zeigen, dass die Recall-Einschätzungen der Probanden nicht exakt sind, aber positiv mit den tatsächlichen Recallwerten korrelieren. Weiterhin ergibt die Studie, dass der Hauptgrund für das Beenden einer Suche das von den Benutzern subjektiv empfundene Gefühl der Informationssättigung ist. Auch Al-Maskari et al. (2006) kommen in diesem Zusammenhang zu dem Schluss, dass die Benutzerwahrnehmung weniger exakt zu sein scheint (vgl. Abschn. 3.3.2.3): „On average, users believed they had completed the task when they had actually achieved 69% recall. In general users were satisfied with the system despite the fact this did not reflect on their performance.“ (ebd., S. 3) Interessanterweise bewegen sich die Recall-Einschätzungen der meisten Teilnehmer in der Studie von Dostert und Kelly (2009) mit 51 bis 60% auf einem ähnlichen Niveau wie in der Studie von Al-Maskari et al. (2006). Dasselbe Verhalten kann auch in der in Abschnitt 3.2.1 beschriebenen Studie von Allan et al. (2005, S. 439) beobachtet werden, in der die Teilnehmer im Mittel nicht mehr als 60% der relevanten Antwortfacetten ausfindig machen. Auch Turpin und Hersh (2001, S. 229) beobachten, dass bei einem größeren Angebot an relevanten Dokumenten der Anteil aufgerufener Dokumente sinkt. Während Nutzer des schlechteren Systems im Rahmen der Question-Answering-Studie 30% der relevanten Dokumente der zehn zuoberst angezeigten

Dokumente nicht aufrufen, beträgt der Anteil bei den Nutzern des besseren Systems 55%. Diese Ergebnisse könnten darauf hindeuten, dass auf diesem Recallniveau für viele Nutzer eine Informationssättigung einsetzt, sie ihr Informationsbedürfnis als befriedigt empfinden und somit mit der von ihnen erbrachten Suchleistung zufrieden sind.

Eine aktuelle Studie, die ebenfalls an diesem Punkt ansetzt, ist die Arbeit von Jiang und Allan (2016) (vgl. Abschn. 3.3.2.3). Die Autoren untersuchen den Zusammenhang zwischen drei *Cumulative Gain Maßen*[5] (vgl. Abschn. 4.2.1.1) und dem von den Probanden wahrgenommenen Sucherfolg. Das Benutzererlebnis wird in dieser Studie über zwei Fragen erhoben: Die erste Frage betrifft die Wahrnehmung der eigenen Suchleistung. Dazu bewerten die Probanden auf einer 5-stufigen Skala, wie gut sie ihre eigene Leistung einschätzen. Die zweite Frage berücksichtigt die Aufgabenschwierigkeit, die von den Probanden ebenfalls auf einer 5-stufigen Skala bewertet wird. Es nehmen insgesamt 20 Probanden an der Untersuchung teil. Alle Teilnehmer bearbeiten vier Aufgaben, zu deren Lösung ihnen jeweils 10 Minuten zur Verfügung stehen. Damit liegen der Analyse 80 Suchsitzungen zugrunde. Zwei Ergebnisse dieser Studie sind besonders bemerkenswert. Erstens korreliert die Einschätzung der eigenen Suchleistung am stärksten mit dem niedrigsten innerhalb einer Suchsitzung auftretenden *normalized Discounted Cumulative Gain*[6] (nDCG), während der höchste nDCG-Wert keine bedeutsamen Korrelationen zu den erhobenen Benutzermaßen aufweist. Dies lässt Jiang und Allan (ebd., S. 288) zu dem Schluss gelangen, dass „[...] a few underperforming queries in a session may substantially affect user experience, while it is common to find well-performing queries in any session." Zweitens korreliert auch der nDCG-Wert der letzten Suchanfrage einer Suchsitzung signifikant mit beiden Benutzermaßen. Tatsächlich weist dieser Wert sogar die stärkste Korrelation mit der von den Teilnehmern empfundenen Aufgabenschwierigkeit unter allen betrachteten Systemleistungsmaßen auf. Der nDCG-Wert der ersten Suchanfrage hingegen korreliert kaum mit den Benutzermaßen. Dieses Ergebnis lässt Jiang und Allan (ebd., S. 288) vermuten, dass „[...] failing to formulate effective queries in later stages of a session may be an indicator of task difficulty."

Zusammenfassend zeigt sich also, dass sich Systemunterschiede nicht nur in der Benutzerleistung widerspiegeln, sondern auch aktiv vom Nutzer wahrgenommen werden können. Darüber hinaus bietet der beobachtete Informationssättingungseffekt in Bezug auf den Recall weiteres Erklärungspotential für den Umstand, dass Systemunterschiede am oberen Ende der Leistungsskala nicht unbedingt zu einer Verbesserung der Nutzerleistung führen: Ab einem gewissen Punkt ist bereits eine ausreichende Anzahl relevanter Dokumente vorhanden, um den Informationsbedarf zu decken, so dass die Benutzer keine weiteren relevanten Dokumente mehr benötigen. Die Ergebnisse von Jiang und Allan (ebd.) hingegen machen erneut deutlich, dass die Interaktion mit dem Suchsystem als Prozess analysiert werden sollte, bei dem einzelne Episoden, wie eine einzelne schlechte Suchanfrage, großen Einfluss auf die Gesamtwahrnehmung des Sucherlebnisses ausüben können. Dieser Aspekt wird im Rahmen der Benutzerleistung im nun folgenden Abschnitt weiter vertieft.

[5] Bei den Cumulative Gain Maßen handelt es sich um eine Gruppe von Systemleistungsmaßen, bei denen die Relevanz aller zurückgegebener Dokumente gewichtet mit ihrer Rankingposition aufsummiert wird (vgl. Abschn. 4.2.1.1).

[6] Bei dem Discounted Cumulative Gain (DCG) handelt es sich um ein Systemleistungsmaß, bei dem die Relevanz aller zurückgegebener Dokumente gewichtet mit ihrer Rankingposition aufsummiert wird. Im Fall des normalized Discounted Cumulative Gain wird dieser DCG-Wert darüber hinaus in Bezug auf eine hypothetische, ideale Sortierung der Rankingliste normalisiert (vgl. Abschn. 4.2.1.1).

3.3. Zur Entstehung von Zufriedenheit im Information Retrieval

Die Integration des Nutzers in den Evaluierungsprozess hat zur Folge, dass neben den reinen Suchergebnissen auch der Suchprozess und das Nutzererlebnis in den Fokus rückt. In diesem Zusammenhang kann Nutzerzufriedenheit als eine Verallgemeinerung der Relevanz angesehen werden, die zusätzlich zu der wahrgenommenen Qualität der Suchergebnisse auch andere Faktoren, wie bspw. die Usability des Systems und die Wahrnehmung der eigenen Leistung, berücksichtigt. In diesem Abschnitt werden verschiedene Ansätze vorgestellt, die die Zufriedenheit von Nutzern mit Suchmaschinen bzw. deren Suchergebnissen zu erklären versuchen. Damit wendet sich die vorliegende Arbeit einem Feld zu, für das vor allem die benachbarten Disziplinen (insb. Psychologie u. Marketing) theoretische Konzepte anbieten. Im Wesentlichen begreifen alle diese Ansätze die Entstehung von Zufriedenheit als einen Abgleich zwischen den gewünschten und den wahrgenommenen Gegebenheiten. Der Erwartungshaltung kommt somit also eine entscheidende Rolle im Entstehungsprozess von Zufriedenheit bzw. Unzufriedenheit zu. Nach der Vorstellung dieser allgemeinen theoretischen Ansätze in Abschnitt 3.3.1 werden in Abschnitt 3.3.2 Zufriedenheitstudien vorgestellt, die sich speziell auf den Kontext der IR-Evaluierung beziehen.

3.3.1. Theorien der Kunden- und Benutzerzufriedenheit

Grundsätzlich existieren in der Kundenzufriedenheitsforschung drei Theorien, die einen Erklärungsbeitrag für die Analyse der Entstehungsgründe von Benutzerzufriedenheit mit Suchmaschinen geben können: das Konfirmations-/Diskonfirmations-Paradigma (C/D-Paradigma), die Assimilations-Kontrast-Theorie und die Attributionstheorie. Diese Theorien werden im Folgenden kurz und ohne Hinweis auf die umfangreiche Forschungsgeschichte dargelegt. Eine ausführlichere Darstellung hierzu findet sich z.B bei Kaiser (2005), Richter (2005) und Homburg (2012). Im Rahmen dieser Arbeit liegt der Schwerpunkt stattdessen auf dem Erklärungsbeitrag dieser Theorien für das Konstrukt der Benutzerzufriedenheit im Bereich der Informationssuche. In Abschnitt 3.3.2 werden deshalb relevante Untersuchungsergebnisse mit vergleichbarer Fragestellung aus dem IR-Kontext präsentiert.

3.3.1.1. Benutzerzufriedenheit nach dem Konfirmations-/Diskonfirmations-Paradigma

In der Kundenzufriedenheitsforschung wird das Entstehen von Zufriedenheit bzw. Unzufriedenheit als Abgleich zwischen den Erwartungen des Kunden und der erlebten Leistung begriffen. Es handelt sich somit um einen Soll-Ist-Vergleich, an dessen Ende das Zufriedenheitsurteil steht. In Anlehnung an die Tatsache, dass die Erwartungen des Kunden entweder erfüllt, enttäuscht oder übertroffen werden können, wird dieser Erklärungsansatz als Konfirmations-/Diskonfirmations-Paradigma bezeichnet. Zahlreiche Studien aus dem Bereich der Marketingforschung können Konfirmation und Diskonfirmation von Kundenerwartungen als wesentliche Einflussgrößen auf die Kundenzufriedenheit identifizieren (Cadotte et al., 1978; Churchill u. Surprenant, 1982). Das C/D-Paradigma bildet darüber hinaus sowohl Ausgangspunkt als auch integrativen Rahmen für speziellere Theorien wie die in den nächsten beiden Abschnitten vorgestellte Assimilations-Kontrast-Theorie und die Attributionstheorie (Homburg u. Stock-Homburg, 2012, S. 20).

Im Folgenden wird der dem C/D-Paradigma zugrunde liegende Vergleichsprozess anhand der schematischen Darstellung in Abbildung 3.1 genauer erläutert. Ausgehend vom Vergleich zwischen wahrgenommener Ist- und erwarteter Soll-Leistung können die folgenden drei Fälle

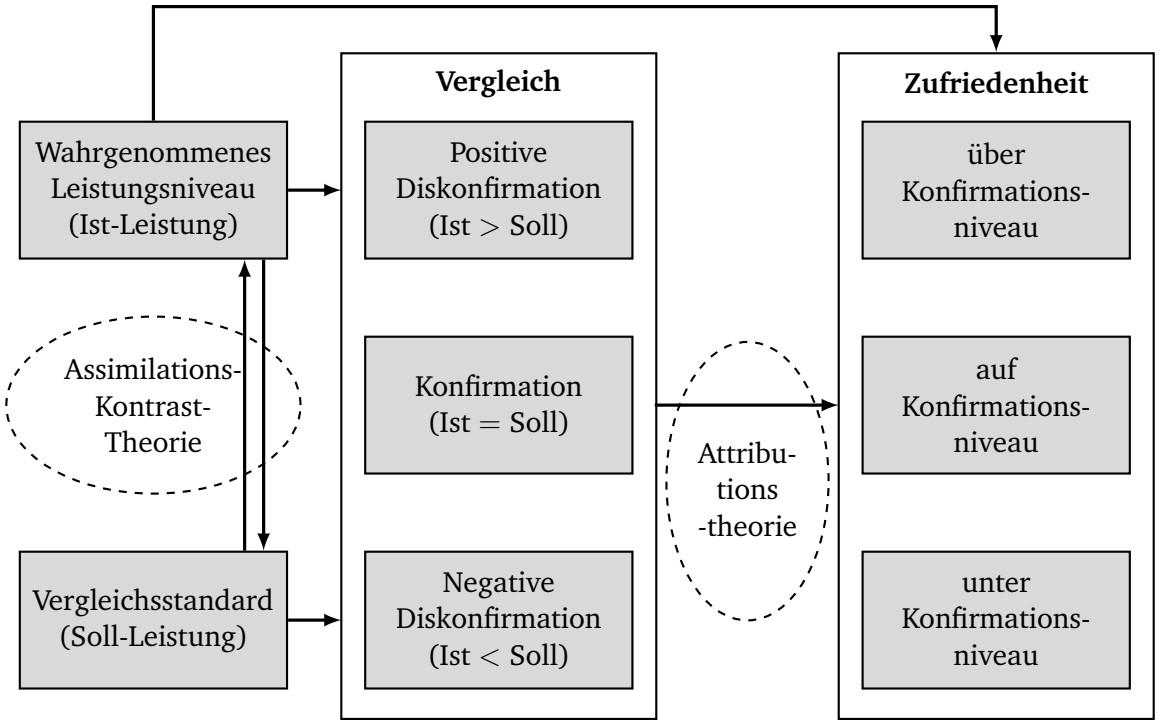


Abb. 3.1.: Schematische Darstellung des Konfirmations-/Diskonfirmationsparadigmas (nach Homburg und Stock-Homburg, 2012, S. 21). Abhängig vom Ergebnis des Soll/Ist-Vergleichs zwischen erwarteter Leistung (Soll-Leistung) und wahrgenommener Leistung (Ist-Leistung) entsteht Zufriedenheit oder Unzufriedenheit. Die Assimilations-Kontrast-Theorie beschreibt mögliche Anpassungen des wahrgenommenen Unterschieds vor dem Soll/Ist-Vergleich, während die Attributionstheorie Anpassungen nach diesem Vergleich erklärt (vgl. Abschn. 3.3.1.2 u. 3.3.1.3).

auftreten: Im einfachsten Fall entspricht die Ist-Leistung gerade der Soll-Leistung, was zu einer Zufriedenheit auf dem sog. Konfirmationsniveau führt. Übertrifft die erlebte Leistung hingegen die antizipierte, resultiert dies in positiver Diskonfirmation, was zu Zufriedenheit über dem Konfirmationsniveau führt. Bleibt die Ist-Leistung hingegen hinter den Erwartungen zurück, erlebt der Kunde negative Diskonfirmation, die zu Unzufriedenheit führt. Insbesondere kann das Übertreffen einer niedrigen Erwartung potentiell zu einer höheren Zufriedenheit führen als das Erfüllen einer hohen Erwartung.

Neben Studien, die die Gültigkeit bzw. Anwendbarkeit des C/D-Paradigmas nachweisen, gibt es jedoch auch Kritikpunkte. Zum einen wird bemängelt, dass das Modell das Vorhandensein von Erwartungen als gegeben voraussetzt. Es kann also nicht auf den Fall angewendet werden, bei dem ein Kunde bspw. zu Beginn keine Erwartungen an das Produkt hatte, aber am Ende trotzdem unzufrieden damit ist (Homburg u. Rudolph, 1997, S. 39). Darüber hinaus ist zu beachten, welche Konzeptualisierung von Erwartungen von den Kunden als Vergleichsstandard herangezogen wird. Wie in Abschnitt 2.1.1.1 dargelegt, stellt sich insbesondere die Frage, ob prädiktive oder normative Erwartungen die Hauptrolle spielen – ob also die wahrgenommene Leistung im Sinne eines „so-soll-es-sein“ oder eines „so-wird-es-sein“ verglichen wird (Bunse, 2000, S. 20). Johnson et al. (1995, S. 700) weisen in diesem Zusammenhang darauf hin, dass sich Erwartungen mit der Zeit an das Marktumfeld anpassen. Es könnte also eine dynamische Komponente geben, bei der Erfahrungen die prädiktive Natur der Erwartung in den Vordergrund

treten lassen.

Darüber hinaus können weitere Effekte sowohl den wahrgenommenen Unterschied zwischen Soll- und Ist-Leistung als auch das aus dem Vergleich resultierende Zufriedenheitsempfinden beeinflussen. Die für den Suchprozess relevanten Mechanismen werden in den folgenden beiden Abschnitten dargestellt.

3.3.1.2. Benutzerzufriedenheit nach der Assimilations-Kontrast-Theorie

Das C/D-Paradigma beruht wesentlich auf dem Vergleich zwischen wahrgenommener und erwarteter Leistung. Mit dem Ziel, das menschliche Verhalten im Fall einer Erwartungsdiskonfirmation genauer zu beschreiben, setzt die Assimilations-Kontrast-Theorie an eben dieser Stelle an und identifiziert Umstände, unter denen eine Verstärkung bzw. Verminderung der wahrgenommenen Unterschiede auftreten kann. Dabei stellt die von Hovland et al. (1957) entwickelte Assimilations-Kontrast-Theorie im Wesentlichen eine Kombination aus zwei weiteren Theorien dar, der Assimilations- und der Kontrasttheorie, die im Folgenden kurz erläutert werden.

Die Assimilationstheorie kann als Erweiterung von Festingers Theorie der kognitiven Dissonanz (Festinger, 1978) angesehen werden (Homburg u. Stock-Homburg, 2012, S. 24). Diese besagt, dass man im Allgemeinen bestrebt ist, einen Zustand kognitiver Konsistenz zu erreichen, d.h. Diskrepanzen zwischen Erwartung und Wahrnehmung durch eine Anpassungsleistung aneinander anzugleichen. Angewendet auf den Kontext des C/D-Paradigmas geht die Assimilationstheorie davon aus, dass Kunden dazu neigen, kognitive Dissonanz vermeiden zu wollen, indem sie Diskrepanzen zwischen wahrgenommener und erwarteter Leistung reduzieren. Dies kann entweder durch eine nachträgliche Korrektur der eigenen Erwartungshaltung oder der Leistungswahrnehmung geschehen. Die Kontrasttheorie hingegen beschreibt den umgekehrten Effekt. Anstatt einen vorhandenen Unterschied zwischen Ist- und Soll-Leistung zu vermindern, führt hier die negative bzw. positive Diskonfirmation zu einer Verstärkung des wahrgenommenen Unterschiedes zwischen Soll- und Ist-Leistung. Im Ergebnis fällt das Zufriedenheitsurteil also besonders positiv oder negativ aus.

Die Assimilations-Kontrast-Theorie führt diese beiden Erklärungsmodelle zusammen. Dazu werden die wahrgenommenen Unterschiede zwischen Soll- und Ist-Leistung in drei Zonen eingeteilt: einen Akzeptanz-, einen Neutralitäts- und einen Ablehnungsbereich (Kaiser, 2005, S. 63; Homburg und Stock-Homburg, 2012, S. 27 f.). Im Akzeptanzbereich, der nur geringe Unterschiede zwischen Soll- und Ist-Leistung umfasst, führt Assimilation zu einer Verringerung der wahrgenommenen Unterschiede. In der Folge bewegt sich die resultierende Zufriedenheit in der Nähe des Konfirmationsniveaus. Fällt der Unterschied hingegen in den Neutralitätsbereich, findet keine Anpassung statt und die Diskrepanz wird im Wesentlichen wahrgenommen, wie sie ist. Im Ablehnungsbereich wiederum tritt ein Kontrasteffekt auf und der wahrgenommene Unterschied fällt nach der Anpassung der Erwartung bzw. des Erlebens größer aus. Dabei ist zu beachten, dass die Breite der drei Bereiche nicht statisch ist, sondern durch situative und individuelle Gegebenheiten beeinflusst werden kann (Kaiser, 2005, S. 63). Damit beschreibt die Assimilations-Kontrast-Theorie in gewisser Weise einen Wechselwirkungseffekt zwischen Erwartungshaltung und wahrgenommener Leistung, bei dem geringe Unterschiede durch den Nutzer unterschätzt, größere Unterschiede hingegen überschätzt werden.

3.3.1.3. Benutzerzufriedenheit nach der Attributionstheorie

Die Attributionstheorie schließlich setzt im Anschluss an den Vergleichsprozess zwischen erwarteter und wahrgenommener Leistung an und zieht für das Zufriedenheitsurteil zusätzlich die Ursachenzuschreibung für die wahrgenommene Leistung in Betracht. Es geht also nicht nur darum, ob die erwartete Leistung erbracht wird, sondern auch warum bzw. warum nicht. „Kunden möchten nicht nur feststellen, ob ein Produkt oder eine Dienstleistung 'funktioniert' oder nicht, sondern wollen (besonders im negativen Fall) auch die Ursache hierfür wissen." (Bunse, 2000, S. 17). Ursprünglich geht die Attributionstheorie auf Forschungen von Heider (1958) und Weiner (1985) im Bereich der Motivationspsychologie zurück. Die Hauptthese besagt, dass Menschen bestrebt sind, Ursachen bzw. Verursacher für die Ereignisse und das Verhalten von Mitmenschen zu finden und dass sie ihr Handeln von der erkannten Ursache abhängig machen. Weiner (ebd.) geht davon aus, dass diese Kausalzuschreibung ein Konstrukt mit den drei Dimensionen Ort, Stabilität und Kontrolle darstellt (Bunse, 2000; Homburg u. Stock-Homburg, 2012). Der Ort der Ursache erfragt dabei, ob die Ursache bei der Person selbst (intern) oder außerhalb (extern) liegt. Die Stabilität der Ursache beschreibt, ob die Ursache als dauerhaft oder flüchtig empfunden wird. Die letzte Dimension schließlich fragt, ob die Ursache als kontrollierbar oder unkontrollierbar wahrgenommen wird.

Es zeigt sich, dass auch das Zufriedenheitsurteil davon abhängen kann, was als Ursache für die Erfüllung bzw. die Enttäuschung der Erwartung angesehen wird. Insbesondere die Tatsache, ob der Kunde sich selbst für den Erfolg bzw. Misserfolg verantwortlich fühlt oder es dem Anbieter zuschreibt, beeinflusst das Zufriedenheitsurteil (Pham et al., 2010). In der Tat führt eine Erwartungserfüllung, die sich der Kunde selbst zuschreibt, zu einem höheren Zufriedenheitsniveau, als wenn der Anbieter dafür verantwortlich gemacht wird (Folkes, 1984). Umgekehrt jedoch fällt die Unzufriedenheit bei einer Enttäuschung der Erwartung geringer aus, wenn sie als selbstverursacht und nicht als durch den Anbieter verantwortet wahrgenommen wird (ebd.). Ein ähnlicher Effekt lässt sich auch in Bezug auf die Stabilität und die Kontrolle beobachten. Die Unzufriedenheit vergrößert sich, falls der Kunde davon ausgehen muss, dass keine Besserung zu erwarten ist, während über eine einmalige Enttäuschung hinweggesehen werden kann. Folkes et al. (1987) weisen dies anhand der Reaktionen auf regelmäßige bzw. überraschende Flugverspätungen nach. Genauso scheinen als außerhalb der Kontrolle des Anbieters wahrgenommene Ursachen zu einer geringeren Unzufriedenheit zu führen (Folkes, 1984). Bunse (2000, S. 18) fasst dies wie folgt zusammen: „Zufriedenheit des Kunden entsteht besonders dann, wenn ein Kunde erfolgreich und aktiv im Austauschprozeß mitwirkt. Unzufriedenheit stellt sich beim Kunden dann ein, wenn stabile externe Ursachen für negative Erfahrungen verantwortlich sind, über die der Anbieter die Kontrolle gehabt hätte."

Die bis hierhin beschriebenen Ansätze zur Entstehung von Zufriedenheit, lassen die Frage außer Acht, wie sich Erwartungen und damit auch die Zufriedenheit im Zeitverlauf verändern (z.B. wenn im Rahmen einer Suchsession eine Reihe unterschiedlicher Suchanfragen gestellt werden). Die folgende Darstellung einer dynamisierten Variante des traditionellen C/D-Paradigmas geht auf diese Fragestellung genauer ein.

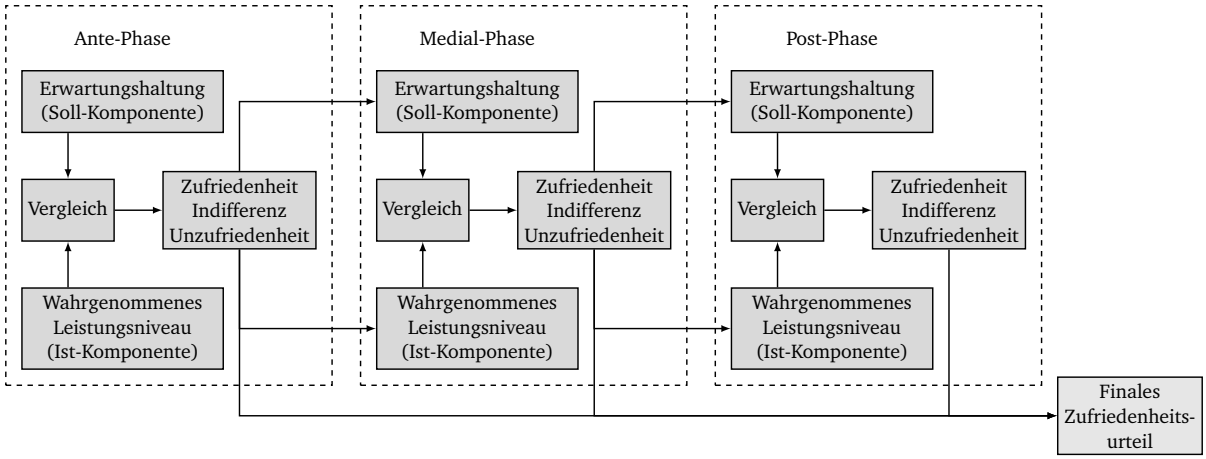


Abb. 3.2.: Schematische Darstellung des dynamisierten Konfirmations-/Diskonfirmations-Paradigmas (nach Richter, 2005, S. 91). Im Falle mehrerer Interaktionen umfasst die Medial-Phase weitere Soll/Ist-Vergleiche zwischen Erwartungshaltung und wahrgenommener Leistung.

3.3.1.4. Dynamisierung des Konfirmations-/Diskonfirmationsparadigmas

In realen Nutzungskontexten basiert die wahrgenommene Leistung häufig nicht auf einer einzelnen zeitlich eng umgrenzten Interaktion des Kunden mit einem Produkt oder einer Dienstleistung. Vielmehr findet eine fortgesetzte Nutzung mit mehreren Interaktionen statt, für die das in Abschnitt 3.1 beschriebene statische C/D-Paradigma nur eingeschränkt Gültigkeit besitzt (Richter, 2005, S. 90; Kaiser, 2005, S. 73). Um diesem stärker prozessorientierten Zufriedenheitsverständnis Rechnung zu tragen, wird das statische C/D-Paradigma um eine dynamische Komponente ergänzt. Grundlage dieses Ansatzes ist das aus der Sozialpsychologie entlehnte Konzept der episodischen Informationsverarbeitung, welches in diesem Zusammenhang besagt, dass unterschiedliche Interaktionen mit einem Produkt als Einzelinformationen wahrgenommen und erinnert und somit auch einzeln in Bezug auf die Zufriedenheit bewertet werden können (Richter, 2005, S. 90). Der Prozess der Zufriedenheitsbildung wird bei diesem Ansatz als lineare Abfolge einzelner Interaktionsepisoden begriffen, die in Folge eines C/D-typischen Soll-Ist-Vergleiches sowohl eine Zufriedenheitsreaktion als auch eine Anpassung der Erwartungshaltung hervorrufen können.

Schematisch ist dieser Prozess in Abbildung 3.2 dargestellt. In der als Antephase bezeichneten ersten Interaktion mit dem Produkt beruht der Soll/Ist-Vergleich auf der initialen Erwartungshaltung des Kunden, welcher zu einem ersten Zufriedenheitsurteil führt. Gleichzeitig erlaubt das Modell eine Anpassung der initialen Erwartungshaltung. Die folgenden, in der Medialphase zusammengefassten Interaktionen, beruhen jeweils auf der aktuell wahrgenommen Leistung sowie der zu diesem Zeitpunkt vorliegenden Erwartungshaltung, die wiederum durch die Nutzungserfahrung beeinflusst werden kann. Auf die die finale Interaktion umfassende Postphase folgt schließlich das Gesamtzufriedenheitsurteil.

Das dynamisierte C/D-Paradigma begreift das finale Zufriedenheitsurteil also als Ergebnis des gesamten Nutzungsprozesses, der alle gesammelten Erfahrungen, die ein Kunde im Verlauf seiner Interaktion mit einem Produkt macht, mit einbezieht. Allerdings macht das Modell keine Aussage über die genauen Wirkungszusammenhänge zwischen den intermediären und dem

abschließenden Zufriedenheitsurteil und deren Einfluss auf die dynamische Entwicklung der Erwartungshaltung (Kaiser, 2005). Tatsächlich nennt Kaiser (ebd., S. 80 f.) mehrere plausible aber gegensätzliche Hypothesen, wie das Zufriedenheitsurteil nach einer Interaktion den folgenden Soll/Ist-Vergleich über einen sog. Ausstrahlungs- bzw. Halo-Effekt beeinflussen kann. Einerseits ist es denkbar, dass sich ein vorläufiges Zufriedenheitsurteil größtenteils auf die wahrgenommene Leistung auswirkt, die Erwartungshaltung jedoch im wesentlichen stabil bleibt. Unter dieser Annahme würde bspw. eine zuvor erlebte Unzufriedenheit zu einer geringeren wahrgenommenen Leistung bei späteren Interaktionen mit einem Produkt führen, als wenn keine negativen Erfahrungen gemacht worden wären. Wirkt sich das vorläufige Zufriedenheitsurteil hingegen stärker auf die Erwartungshaltung des Kunden aus, ist mit einem gegensätzlichen Verhalten zu rechnen. Die erlebte Unzufriedenheit führt zu einer geringeren Erwartungshaltung, was bei der darauf folgenden Episode bei gleicher wahrgenommener Leistung zu einem positiveren Zufriedenheitsurteil führt. Schließlich ist es auch vorstellbar, dass sowohl die wahrgenommene Leistung als auch die Erwartungshaltung durch die vorgelagerte Interaktion verändert werden und somit die Richtung des nachfolgenden Zufriedenheitsurteils von der relativen Stärke dieser beiden Effekte abhängig ist (ebd., S. 81).

Genaueren Aufschluss über die Dynamik der Erwartungshaltung im Kontext der Kundenzufriedenheit mit Dienstleistungen gibt ein von Boulding et al. (1993) entwickeltes dynamisches Strukturmodell, das es erlaubt den Einfluss normativer und prädiktiver Erwartungen (vgl. Abschn. 2.1.1.1) auf das Zufriedenheitsurteil sowie ihre zeitliche Entwicklung zu betrachten. Eine von Boulding et al. (ebd.) durchgeführte Studie legt nahe, dass sich normative Erwartungen als relativ stabil in Bezug auf die Interaktion mit einem Produkt verhalten und im Wesentlichen nur durch ihre Übererfüllung zu verändern sind. Prädiktive Erwartungen hingegen scheinen schneller an die wahrgenommene Leistung angepasst zu werden. Auch in Bezug auf den Einfluss auf nachgelagerte Zufriedenheitsurteile unterscheiden sich normative und prädiktive Erwartungen. Während höhere prädiktive Erwartungen zu einer geringeren wahrgenommenen Leistung führen, wirken sich hohe normative Erwartungen positiv auf das Zufriedenheitsurteil aus (ebd.).

In Bezug auf eine prozessorientierte Sichtweise auf die Informationssuche bietet das dynamisierte C/D-Paradigma einen plausiblen theoretischen Rahmen, um die Dynamik von Erwartungshaltung und wahrgenommener Systemqualität zu evaluieren. Allerdings muss zunächst in Anlehnung an die vorgestellten möglichen Wirkungszusammenhänge überprüft werden, welche Hypothese im Kontext der Informationssuche zutreffend ist. Dies ist Gegenstand und eine der Hauptforschungsfragen der dritten Nutzerstudie im praktischen Teil dieser Arbeit (vgl. Kap. 7).

3.3.2. Zufriedenheit im Kontext der Informationssuche

In diesem Abschnitt wird der aktuelle Stand der Forschung im Bereich der Benutzerzufriedenheit mit Informations- und Suchsystemen dargestellt. Der Abschnitt ist in vier Unterabschnitte gegliedert, die sich jeweils mit einem spezifischen Aspekt der Benutzerzufriedenheit im Kontext der Informationssuche befassen. Im ersten Teil werden Studien besprochen, die den Einfluss von Erwartungen auf die Zufriedenheit von Benutzern untersuchen. Im Hinblick auf die theoretische Fundierung des Untersuchungsgegenstands ist es das Ziel dieser Literaturübersicht, die Relevanz von Benutzererwartungen bei der Informationssuche zu verdeutlichen und die Erklärungspotenziale und Grenzen der zuvor vorgestellten Ansätze für den Bereich der Suche herauszuarbeiten.

In den folgenden beiden Abschnitten werden Studien besprochen, die sich mit dem Einfluss von Systemqualität und Sucherfolg auf die Benutzerzufriedenheit auseinandersetzen. Beide Aspekte sind kaum voneinander zu trennen, nicht zuletzt, weil die Systemqualität im IR-Kontext durch die Interaktion mit dem System wahrgenommen wird und der Sucherfolg zu einem gewissen Grad von der Systemqualität abhängt. Im Rahmen der hier vorgestellten Untersuchungsbeispiele werden dennoch beide Einflussfaktoren zunächst getrennt betrachtet, um die spezifischen Anforderungen und daraus resultierenden Anpassungsstrategien besser beurteilen zu können. Der vierte Abschnitt behandelt schließlich dynamische Aspekte des Zufriedenheitsurteils in IIR-Studien.

3.3.2.1. Der Einfluss von Erwartungen auf die Zufriedenheit

Wie in den einleitenden Abschnitten dieses Kapitels dargelegt (vgl. Abschn. 3.3.1), stellen im Rahmen der Kundenzufriedenheitsforschung Erwartungen neben der wahrgenommenen Leistung den zentralen Einflussfaktor auf die Zufriedenheit dar. Die Vermutung, dass Erwartungen auch im Kontext der Informationssuche eine entscheidende Rolle spielen, verdeutlicht das folgende Zitat von Cox und Fisher (2004, S. 1): „When searching for a lost object, casual observation suggests that we are most satisfied when we find the object when we least expect to do so. That is, our satisfaction is mediated by our expectation of success." Im Folgenden werden nun Studien vorgestellt, die diese These und die Wirksamkeit der in Abschnitt 3.3.1 diskutierten Mechanismen der Zufriedenheitsreaktion im Kontext der Informationssuche betrachten.

In den Grenzbereich zwischen IR- und traditioneller Marketingforschung fällt dabei die Studie von Jansen et al. (2007) zur Markenwahrnehmung bei Suchmaschinen (vgl. Abschn. 2.1.1.3). In einem Laborexperiment mit 32 Testpersonen untersuchen sie unter Verwendung eines Within-Subject-Designs den Einfluss der Markenwahrnehmung auf die Relevanzbeurteilung. Dazu bearbeitet jede Testperson vier Suchaufgaben mit vier verschiedenen Suchmaschinen. Diese umfassen neben den drei bekannten Suchmaschinen von Google, MSN und Yahoo auch ein den Teilnehmern unbekanntes System. Die angezeigten Ergebnisdokumente sind für jede Suchmaschine und Aufgabe identisch, einzig die Markenelemente auf der Suchergebnisseite unterscheiden sich. Es zeigt sich, dass Erwartungen in der Tat einen deutlichen Einfluss auf die Qualitätsbewertung der Suchergebnisse haben. So schneidet im Vergleich die den Teilnehmern unbekannte Suchmaschine am schlechtesten ab (ebd.).

In einer frühen Studie analysiert Su (1994) den Zusammenhang zwischen 20 IR-Evaluierungsmaßen und wahrgenommenem Sucherfolg im Rahmen eines durch professionelle Rechercheure unterstützten Suchprozesses mit realen Informationsbedürfnissen der Probanden. Es zeigt sich, dass insbesondere die Precision der Retrievalergebnisse keinen großen Einfluss auf die Zufriedenheit der Teilnehmer mit dem Sucherfolg hat bzw. die Teilnehmer auch bei Ergebnissen mit geringer Precision ihre Zufriedenheit mit dem Sucherfolg zum Ausdruck bringen. Su (ebd.) führt diese Tatsache u.a. auf die Erwartungen der Probanden zurück, da die Teilnehmer in Interviews bestätigen, im Rahmen ihres Informationsbedürfnisses größeren Wert auf die Vollständigkeit, also den Recall, der Suchergebnisse zu legen. Darüber hinaus lassen die Interviewaussagen der Teilnehmer auch Anzeichen von positiver Diskonfirmation für die Precision erkennen, da ihre ursprüngliche Erwartung an die Precision der Ergebnisse übertroffen wird. Ein Teilnehmer berichtet bspw.: „I was having such a difficult time finding anything, such a high proportion (51%)

is definitely satisfactory. If I have 25% I will still be satisfied.“ (Su, 1994, S. 212) Während ein anderer Proband zu Protokoll gibt: „My expectation had been 1 in 7 or 1 in 3 that it might be (relevant). I was pleased that . . . “ (ebd., S. 212). In ähnlicher Weise lassen sich auch Belege für konfirmatorisches Verhalten finden, bei denen die Nutzer Zufriedenheit mit einer geringen Precision ausdrücken, da sie diese erwartet hätten.

Obwohl diese Forschungsarbeiten einen klaren Zusammenhang zwischen Benutzererwartungen und -zufriedenheit aufzeigen, finden die Erklärungspotentiale der in Abschnitt 3.3.1 vorgestellten Theorien bislang kaum Berücksichtigung bei der Analyse von Suchprozessen. Im Folgenden wird der Rahmen daher etwas weiter gespannt und auch auf Studien ausgedehnt, die sich mit der Nutzung von betrieblichen Informationssystemen auseinandersetzen, deren Nutzung von Außen vorgegeben ist.

Kwahk und Oh (2009) bspw. führen eine Fragebogenstudie mit 208 Nutzern eines verpflichtend eingeführten Enterprise-Resource-Planning (ERP)-Systems durch, um u.a. den Einfluss von Ergebniserwartungen auf die Nutzerzufriedenheit zu untersuchen. Es werden zwei Arten von Ergebniserwartung erfasst: Nützlichkeit des Systems bei der Lösung von Aufgaben (Leistungssteigerung) und Auswirkung der Systemnutzung auf die eigene Reputation im Unternehmen. Während sich der erste Aspekt also auf die erwartete Systemleistung bezieht, betont der zweite Faktor den persönlichen Nutzen, den die Teilnehmer für sich selbst erwarten. Neben der Benutzerzufriedenheit werden auch die wahrgenommene System- und Informationsqualität erfasst. Kwahk und Oh (ebd.) können nachweisen, dass Erwartungen positiv mit der wahrgenommenen Systemqualität korrelieren, was sie als möglichen Effekt der kognitiven Dissonanz-Reduktion interpretieren (vgl. Abschn. 3.3.1.2). Die wahrgenommene Systemqualität wiederum zeigt einen signifikanten Einfluss auf die Nutzerzufriedenheit und kann $61{,}8\%$ der beobachteten Varianz erklären.

In einer dreiteiligen auf simulierten Arbeitsaufgaben (vgl. Borlund und Ingwersen (1997), Abschn. 4.1.3.2) beruhenden Studie zur Verteilung eines Werbebudgets mit Hilfe eines Informationssystems analysieren Szajna und Scamell (1993) den Zusammenhang zwischen Erwartung, Zufriedenheit und Benutzerleistung. Als Teilnehmer werden 159 Bachelorstudenten der Betriebswirtschaftslehre rekrutiert, die einer von drei Untersuchungsgruppen (hohe, mittlere und niedrige Erwartung) zugeteilt werden. Während kein Einfluss der Erwartungsmanipulation auf die Benutzerleistung nachgewiesen werden kann, zeigt sich erneut eine positive Korrelation mit der Benutzerzufriedenheit. Die Tatsache, dass bei Nutzung desselben Systems unterschiedliche Erwartungshaltungen zu unterschiedlichen Zufriedenheitsreaktionen führen, wird wiederum als Hinweis auf einen Dissonanzeffekt angesehen. Auf dynamische Aspekte der Studie wird in Abschnitt 3.3.2.4 genauer eingegangen.

Zurückkommend auf den speziellen Kontext der IR-Evaluation soll an dieser Stelle auch noch einmal auf eine Studie von Cox und Fisher (2004) eingegangen werden, die bereits in Abschnitt 2.1.1.3 im Rahmen von Erwartung als Einflussfaktor im IR-Kontext vorgestellt wird. Ähnlich wie bei Kwahk und Oh (2009) wird in dieser Studie der Einfluss von Erfolgserwartungen und wahrgenommener Systemqualität auf die Nutzerzufriedenheit analysiert. Die Erfolgserwartung wird hierbei direkt über die Aufgabenstellung operationalisiert, indem die Suchaufgaben so konstruiert sind, dass sie in zwei Fällen hohe und in zwei Fällen niedrige Erwartungen hervorrufen

sollen (vgl. Abschn. 2.1.1.3). Das Testdesign umfasst keine interaktive Komponente mit dem Testsystem, da die Probanden sowohl die Aufgabenstellung als auch die zu verwendete Suchanfrage vorgegeben bekommen. Im Rahmen der Erwartungsmessung werden deshalb sowohl die Aufgabenschwierigkeit an sich, als auch die speziellen Erfolgsaussichten in Bezug auf die vorgegebene Suchanfrage erfasst. Im Sinne des in Abschnitt 3.3.1.1 beschriebenen Vergleichsprozesses zwischen erwarteter und wahrgenommener Leistung verwenden Cox und Fisher (2004) als Prädiktor für die Benutzerzufriedenheit die Differenz zwischen gemessener Erwartung und wahrgenommener Systemqualität. Für alle vier Suchaufgaben lässt sich ein signifikanter Zusammenhang zwischen dieser *Reaktion/Erwartungs*-Differenz und der Nutzerzufriedenheit nachweisen. Die Berücksichtigung sowohl der Aufgabenschwierigkeit als auch der konkreten Suchanfrage macht deutlich, dass Cox und Fisher (ebd.) von einem Erwartungsbegriff ausgehen, der im Rahmen eines interaktiven IR-Experiments auch den Eigenbeitrag des Nutzers zum Sucherfolg berücksichtigen würde. Auf diese Beteiligung des Nutzers in den Suchprozess wird im Folgenden genauer eingegangen.

In der Tat fügt die Partizipation des Nutzers dem in Abschnitt 3.3.1.1 dargestellten Soll-Ist-Vergleich zwischen erwarteter und wahrgenommener Leistung einen neuen Aspekt hinzu. Da das Suchergebnis im Rahmen der Interaktion zwischen Nutzer und System zustande kommt, wird nicht nur das Ergebnis, sondern auch sein Entstehungsprozess wahrgenommen und bewertet. Der Nutzer ist also aktiv am Retrievalprozess beteiligt und nicht passiver Rezipient. Die Bewertung des Retrievalergebnisses ist somit zumindest teilweise auch immer eine Bewertung des eigenen Beitrags. Interessant ist in diesem Zusammenhang, dass die Beurteilung des Sucherfolgs durch die Nutzer selbst häufig positiver ausfällt als eine entsprechende Fremdbeurteilung (Kelly et al., 2008b; Lewandowski, 2014). Als Erklärung für dieses Verhalten lassen sich unterschiedliche Lesarten finden (Kelly et al., 2008b). Möglicherweise sind Nutzer von Suchmaschinen nicht hinreichend in der Lage, ihre eigene Suchleistung zutreffend einzuschätzen oder sie benötigen eine längere Trainingsphase, um sich an das System zu gewöhnen. Zudem können testspezifische Faktoren wie die soziale Erwünschtheit oder der soziale Druck durch andere (wenn mehrere Probanden gleichzeitig teilnehmen) eine Rolle spielen (vgl. Abschn. 4.2.3.2). Zu einem ähnlichen Schluss kommt auch Kelly et al. (2008a, S. 124 f.): „Subjects may view the success or failure of a system as a reflection of their own abilities rather than as a reflection of the system's abilities. For instance, subjects might believe that responding negatively to questions such as how easy a system was to learn to use, how easy a system was to use, or how satisfied they were with their performances, reflects on them rather than the system. When people are part of the process as they are in interactive IR, they may feel that they are at least in part responsible for a system's performance. Thus, people may view negative ratings of systems as negative ratings of themselves, and avoid using such ratings." Im Sinne der in Abschnitt 3.3.1.3 diskutierten Attributionstheorie kann diese höhere Zufriedenheitbewertung darüber hinaus auch durch die Ortsdimension erklärt werden, die durch die Beteiligung des Nutzers am Suchprozess immer einen internen also selbstbezogenen Anteil besitzt. Dieser Zusammenhang wird auch von Kelly et al. (2008b) in einer bereits in Abschnitt 2.2.2 vorgestellten Untersuchung vermutet. In derselben Studie findet sich ein weiterer Hinweis auf sozial erwünschtes Verhalten: Ein realistisches Feedback zur eigenen Suchleistung führt auch zu einer nachträglichen Zufriedenheitskorrektur bei

Items, die sich auf die Benutzeroberfläche des verwendeten Suchsystems beziehen. Im Gegensatz zu der durch die Teilnehmer individuell erbrachte Suchleistung bleibt dieses Interface jedoch während des Experiments unverändert und sollte deshalb Zufriedenheitsreaktionen hervorrufen, die unabhängig vom Sucherfolg sind. Welche der diskutierten Erklärungsmodelle letztlich für die beobachtete Zufriedenheitsanpassung verantwortlich sind oder ob es sich um ein Zusammenspiel der verschiedenen Effekte handelt, kann an dieser Stelle allerdings nicht abschließend beantwortet werden.

In der Gesamtschau lassen die dargestellten Studien den Schluss zu, dass Benutzererwartungen einen wesentlichen Beitrag zur Zufriedenheit leisten. Darüber hinaus kann gezeigt werden, dass die in Abschnitt 3.3.1 diskutierten Mechanismen auch im Suchkontext als Erklärungsmodelle für die Entstehung von Nutzerzufriedenheit dienen können. Insbesondere bei Studien der Informationssystemforschung zur Systemnutzung in stärker verpflichtenden Kontexten lassen sich auch Effekte der kognitiven Dissonanz beobachten. Es ist jedoch unklar, inwieweit sich diese auf den Kontext der Internetsuche übertragen lassen. So verringert die Verfügbarkeit alternativer Suchmaschinen die Notwendigkeit, ein gegebenes System trotz negativer Nutzungserfahrung weiter zu verwenden. Die Dissonanztheorie wird jedoch vor allem zur Erklärung einstellungsdiskrepanten Handelns, in diesem Fall also das Weiterverwenden einer unbefriedigenden Suchmaschine, herangezogen. Interessant erscheint in diesem Zusammenhang allerdings die Frage, inwieweit die Vorgabe eines Suchsystems in interaktiven IR-Experimenten die im Kontext der ERP-Systeme beobachteten Dissonanzeffekte als untersuchungsbedingte Störfaktoren reproduziert. Ein weiteres zentrales Ergebnis dieses Abschnitts stellt der Einfluss der Eigenbeteiligung des Nutzers am Retrievalergebnis auf das subjektive Zufriedenheitsempfinden dar. Die Interaktion mit dem Suchsystem fügt der Zufriedenheitsbildung somit eine weitere Komponente hinzu, die wiederum von der wahrgenommenen Systemqualität und dem Sucherfolg abhängig ist. Im Fokus der nun folgenden beiden Abschnitte stehen aus diesem Grund Studien, die den Zusammenhang zwischen dem Erfolg eines Systems oder einer Suche und der Benutzerzufriedenheit zum Thema haben.

3.3.2.2. Vergleiche zwischen Systemqualität und Zufriedenheit

Wie in Abschnitt 3.3.1.1 dargelegt, wird die Entstehung von Zufriedenheit als Soll-Ist-Vergleich zwischen Erwartungen und wahrgenommener Leistung begriffen. Nachdem im vorangegangenen Abschnitt die Rolle der Nutzererwartung in diesem Prozess analysiert wurde, werden nun Studien betrachtet, die den Einfluss der Systemleistung untersuchen. Bevor im nächsten Abschnitt explizit auf die wahrgenommene Leistung bzw. den individuellen Sucherfolg eingegangen wird, stellt dieser Abschnitt die Frage, inwieweit objektive Effektivitätsmaße, die nicht das individuelle Qualitätsempfinden der Nutzer berücksichtigen, bereits als Indikatoren für die Nutzerzufriedenheit dienen können. Ziel ist es somit, zu untersuchen, welche Effektivitätsmaße aus der systemorientierten IR-Evaluierung, wie bspw. Recall und Precision, gute Prädiktoren für die Nutzerzufriedenheit darstellen. Im Spannungsfeld zwischen system- und nutzerzentrierter Evaluierung betrifft dies also die generelle Frage, ob systemorientierte Ergebnisse auf den Nutzerkontext generalisierbar sind. Trotz der Relevanz dieser Fragestellung finden sich erst in jüngerer Zeit vermehrt Studien zu diesem Thema: „[...] given their importance in IR evaluation, one might assume that the relationship between user satisfaction and, say, average precision has been

thoroughly studied and is well understood. Unfortunately, this is not the case. User studies trying to find correlations between user satisfaction and various effectiveness measures are a relatively recent phenomenon."(Büttcher et al., 2010, S. 410)

Der Großteil der Studien betrachtet den Einfluss von Recall und Precision eines Suchsystems auf die Zufriedenheit der Nutzer (Su, 1994; Su, 2003; Thomas u. Hawking, 2006; Kelly et al., 2007; Al-Maskari u. Sanderson, 2010). So zeigt sich in der bereits im vorangegangenen Abschnitt beschriebenen Studie von Su (1994), dass die Precision der Ergebnisse im Gegensatz zum Recall nur eine untergeordnete Rolle in Bezug auf die Nutzerzufriedenheit spielt. Insbesondere räumen Teilnehmer dieser Studie dem Recall eine höhere Relevanz ein: Auf einer Wichtigkeitsskala von 1 (gering) bis 7 (hoch) ermittelt Su (ebd., S. 213) für die Precision ein mittleren Wert von $4{,}7$, während der Recall einen Wert von $6{,}2$ erreicht. Wie in Abschnitt 3.3.2.1 bereits angedeutet, trägt auch die Erwartungshaltung der Nutzer dazu bei, dass eine hohe Precision aus Benutzersicht nicht in jedem Fall mit Sucherfolg verknüpft ist. So kann eine Suche mit einem hohen Precisionwert bspw. viele schon bekannte Referenzen enthalten (ebd., S. 216). Su (ebd.) schlussfolgert deshalb, dass die Zufriedenheit mit der Precision der Suche einen besseren Indikator für Sucherfolg darstellt als der objektiv erzielte Precisionwert. Im Rahmen einer Folgestudie relativiert Su (2003, S. 1220) jedoch die Aussage, dass Precision kein guter Indikator für Sucherfolg sei insofern, dass die Bevorzugung von Recall oder Precision von der jeweiligen Zielgruppe abhängt. In dieser zweiten Studie untersucht Su (ebd.) das Such- und Evaluierungsverhalten von 36 Probanden, die im Rahmen eines Within-Subject-Designs vier verschieden Suchmaschinen zur Informationssuche verwenden (Alta Vista, Excite, Infoseek u. Lycos). Wie in der ersten Studie stammen die spezifischen Informationsbedürfnisse auch in dieser Studie von den Probanden selbst, jedoch mit dem Unterschied, dass es sich bei einem Großteil der Teilnehmer der ersten Studie um Doktoranden und wissenschaftliche Mitarbeiter handelt, während in der zweiten Studie Bachelorstudenten angeworben werden. Su (ebd., S. 1220) vermutet, dass die hohe Präferenz für die Vollständigkeit der Suchergebnisse aus dem Umstand resultiert, dass die Informationsbedürfnisse der Teilnehmer der ersten Studie im Zusammenhang mit ihren jeweiligen Dissertationsprojekten und Projektanträgen stehen und es in diesen Fällen besonders wichtig ist keine Referenzen zu übersehen. Die Nutzerzufriedenheit hängt also auch hier vom Kontext der Suche und dem tatsächlichen Informationsbedürfnis ab.

Die Bedeutung der Precision im Kontext von Internetsuchen wird in einer ähnlich angelegten Studie von Johnson et al. (2003) bestätigt. Auch in diesem Fall arbeiten die Probanden mit drei verschiedenen Suchmaschinen (Excite, NorthernLight u. HotBot) und bewerten anschließend die Gesamtleistung sowie ihre Zufriedenheit anhand der Kategorien Effektivität, Effizienz, Nützlichkeit und Interaktion. In diesem Zusammenhang zeigen Effizienz und Effektivität die stärkste Korrelation mit der Bewertung der Gesamtleistung. Darüber hinaus bewerten die Probanden ihre Zufriedenheit mit der Precision der Suchergebnisse, welche stark mit der Bewertung der Effektivität korreliert, was als weiteres Indiz für den Einfluss der Precision auf die Zufriedenheit mit der Systemleistung gewertet werden kann.

Darüber hinaus finden sich in der Literatur Studien, in denen die Systemleistung experimentell kontrolliert und aktiv manipuliert wird, um eine genauere und vergleichende Beurteilung des Systemeinflusses zu ermöglichen (Thomas u. Hawking, 2006; Kelly et al., 2007; Al-Maskari u.

Sanderson, 2010). In der bereits in Abschnitt 3.2.2 im Kontext des wahrgenommenen Sucherfolgs vorgestellten Studie von Thomas und Hawking (2006) werden vier Experimente zur Untersuchung der Wahrnehmung von Qualitätsunterschieden von Suchergebnislisten diskutiert. Dabei werden den Probanden für jede Suchaufgabe parallel zwei Ergebnislisten präsentiert, deren relative Systemqualität zu bewerten ist. Die Manipulation der Systemleistung erfolgt durch die Darstellung vorderer bzw. weiter hinten liegender Google-Ergebnisse für die entsprechende Suchanfrage. Die erhobenen Nutzerinteraktionen umfassen die Liste mit dem ersten aufgerufenen Dokument, die Liste mit dem letzten aufgerufenen Dokument, die Liste mit der höchsten Anzahl aufgerufener Dokumente und die Liste mit dem am höchsten gerankten aufgerufenen Dokument. In allen vier Experimenten zeigt sich, dass die Teilnehmer in der Lage sind, den Qualitätsunterschied zwischen den Suchmaschinen wahrzunehmen. Darüber hinaus korreliert das Qualitätsurteil mit der bei den Nutzerinteraktionen präferierten Liste. Insofern können Thomas und Hawking (ebd.) nachweisen, dass sich das Zufriedenheitsurteil der Teilnehmer auch im Nutzerverhalten widerspiegelt. Zu ähnlichen Ergebnissen kommen auch die Studien von Kelly et al. (2007) und Al-Maskari und Sanderson (2010).

Die Nutzerstudie von Kelly et al. (2007) umfasst insgesamt drei Experimente, die bestätigen, dass Ranking und Precision einen signifikanten Einfluss auf die wahrgenommene Qualität der Suchergebnisse ausüben (vgl. Abschn. 3.2.2). Dabei wird in den ersten beiden Experimenten ausschließlich das Ranking, im letzten Experiment hingegen nach Möglichkeit ausschließlich die Precision der Ergebnislisten variiert. Im Anschluss an jede Suchaufgabe evaluieren die Probanden die Leistung des Suchsystems. Da jeder Teilnehmer Aufgaben mit allen Systemqualitäten bearbeitet, besteht darüber hinaus die Möglichkeit, das Qualitätsurteil für die zuvor verwendeten Suchsysteme zu korrigieren, um eine relative Bewertung zu ermöglichen. In Übereinstimmung mit der Studie von Thomas und Hawking (2006) können auch Kelly et al. (2007) nachweisen, dass die Probanden in der Lage sind, die Qualitätsunterschiede zwischen den Suchmaschinen wahrzunehmen. In Bezug auf Ranking und Precision kann darüber hinaus gezeigt werden, dass die Precision der Ergebnislisten einen stärkeren Effekt auf die Leistungsbeurteilung ausübt als das Ranking der Dokumente. Al-Maskari und Sanderson (2010) schließlich untersuchen den Einfluss von Systemleistung, Benutzerleistung, Aufwand sowie individueller Voraussetzungen der Suchenden auf die Benutzerzufriedenheit. Auch diese Untersuchung ist als Within-Subject-Design angelegt. 56 Probanden bearbeiten dabei acht Aufgaben, je vier mit einem besseren und vier mit einem schlechteren System, wobei den Probanden der Systemunterschied nicht bekannt ist. Als Testkorpus dienen in dieser Studie 56 Suchthemen der TREC-8 Dokumentenkollektion. Zur Manipulation der Systemleistung vergleichen die Autoren die AvP-Werte von drei IR-Systemen und wählen für jedes Suchthema die beiden Systeme aus, die sich am stärksten unterscheiden. Neben einer 3-stufigen Relevanzbewertung der Dokumente beurteilen die Testpersonen außerdem auf einer 4-stufigen Skala ihre Zufriedenheit mit den Suchergebnissen. Auch in dieser Studie ist der Zufriedenheitsgrad der Nutzer des besseren Systems signifikant höher als jener der Nutzer des schlechteren Systems.

Zum Abschluss dieses Abschnitts werden zwei Studien beschrieben, die die Nutzerzufriedenheit mit den im Rahmen der Sucherfolgswahrnehmung eingeführten Cumulative Gain Maßen kontrastieren (vgl. Abschn. 3.2.2). Diese Maße bewerten sowohl die Relevanz der zurückge-

lieferten Dokumente als auch die Position relevanter Dokumente in der Ergebnisliste. Genauer wird die Relevanz eines Dokuments als geringer bewertet, je weiter hinten in der Rankingliste es erscheint, indem sein Relevanzwert durch eine Funktion der Rankingposition (typischerweise der Logarithmus) geteilt wird (vgl. Abschn. 4.2.1.1).

Huffman und Hochster (2007) führen eine Studie auf Grundlage 200 realer Google-Anfragen aus dem Jahr 2006 durch. Zunächst klassifizieren Juroren die Suchanfrage als navigational, transactional oder informational, bewerten die Relevanz der ersten drei von Google zurückgelieferten Ergebnisse und formulieren ein der Suchanfrage zugrunde liegendes Informationsbedürfnis. Im zweiten Schritt verwendet eine zweite Gruppe von Probanden diese vorgegebenen Suchanfragen als Ausgangspunkt einer Google-Recherche zu diesen Informationsbedürfnissen und bewertet abschließend ihre Zufriedenheit mit der Suche. Es zeigt sich, dass ein einfaches Cumulative Gain Maß, das die Relevanz der ersten drei Ergebnisse gewichtet mit ihrer Rankingposition berücksichtigt, bereits eine hohe Korrelation mit dem finalen Zufriedenheitsurteil aufweist. Des Weiteren ist ein Unterschied zwischen navigationalen und nicht-navigationalen Suchen zu beobachten: Während im ersten Fall im Wesentlichen nur die Relevanz des ersten Ergebnisses mit der Zufriedenheit korreliert, bleibt bei nicht-navigationalen Suchen dieser Zusammenhang auch für das zweite und dritte Suchergebnis bestehen.

In einer anderen Google-basierten Nutzerstudie mit 26 Teilnehmern untersuchen Al-Maskari et al. (2007) den Zusammenhang zwischen verschiedenen Systemleistungsmaßen und der Nutzerzufriedenheit. Jeder Teilnehmer bearbeitet vier Suchaufgaben aus einem Pool von insgesamt 104 Informationsbedürfnissen und bewertet die ersten zehn Dokumente der nach ihrer Meinung besten Anfrage anhand einer 3-stufigen Relevanzskala (*highly-relevant*, *reasonably relevant* u. *not relevant*). Des Weiteren bewerten die Probanden ihre Zufriedenheit mit dem Ranking, der Vollständigkeit und der Genauigkeit der Suchergebnisse. Die Studie kann starke Korrelationen zwischen der Nutzerzufriedenheit und der Precision sowie dem Cumulative Gain der Ergebnislisten nachweisen. Darüber hinaus finden sich moderate Korrelationen mit dem DCG. Des Weiteren korrelieren auch die Ergebnisse der drei Zufriedenheitsdimensionen miteinander.

In der Gesamtschau zeigen die diskutierten Studien, dass Nutzer in der Lage sind, Systemunterschiede wahrzunehmen und dass sich Unterschiede in der Systemqualität auch in ihrem Zufriedenheitsurteil widerspiegeln. Insbesondere zeigt sich, dass die Precision der Suchergebnisse in einer Reihe von Experimenten eine nachweisbare Korrelation mit der Zufriedenheit der Teilnehmer aufweist. Aus diesem Grund erscheinen Precision und Average Precision (AvP) geeignete Kandidaten, um im Rahmen der hier durchgeführten Experimente die Systemleistung zu manipulieren. In Abschnitt 3.3.2.1 wird argumentiert, dass durch die Partizipation des Nutzers am Suchprozess die Nutzerzufriedenheit auch immer eine Bewertung der eigenen Leistung enthält. Aus diesem Grund wird im folgenden Abschnitt anstelle des Einflusses der objektiven Systemleistung der Einfluss des individuellen Sucherfolgs auf die Nutzerzufriedenheit diskutiert.

3.3.2.3. Zum Zusammenhang zwischen Sucherfolg und Zufriedenheit

Beim Übergang von systemorientierter zu interaktiver IR-Evaluierung rückt der Nutzer in den Mittelpunkt der Betrachtung. Aus diesem Grund erscheint es natürlich, nicht nur den Zusammenhang zwischen objektiver Systemleistung und Nutzerzufriedenheit zu untersuchen, sondern auch den Einfluss des individuellen Sucherfolgs zu analysieren. Dies erlaubt insbesondere neben der

auf Jurorurteilen basierenden Relevanz von Dokumenten auch die individuelle Relevanzwahrnehmung der Probanden sowie den individuellen Aufwand zur Lösung einer Suchaufgabe mit in die Betrachtung einzubeziehen. Darüber hinaus kann die Wahrnehmung des eigenen Sucherfolgs durch den Nutzer als weitere Komponente der Nutzerzufriedenheit erhoben werden. Der folgende Abschnitt stellt eine Reihe von Studien und ihre unterschiedlichen Ansätze zur Messung der Effizienz und des Sucherfolgs vor. Dabei werden zunächst Studien besprochen, die den Einfluss des Nutzeraufwands auf die Zufriedenheit analysieren, bevor im zweiten Teil Arbeiten vorgestellt werden, die sich mit dem Zusammenhang zum Sucherfolg beschäftigen.

Die bereits im Kontext der Systemleistung in Abschnitt 3.3.2.2 im Detail vorgestellte Nutzerstudie von Al-Maskari und Sanderson (2010) beschäftigt sich auch mit dem Zusammenhang zwischen Aufwand und Nutzerzufriedenheit. Der Nutzeraufwand wird hier anhand von zwei Variablen berücksichtigt. Zum einen wird die Anzahl der gestellten Suchanfragen erhoben. Zum anderen wird für jede Testperson die Rankingposition, des in der Suchergebnisliste am höchsten gerankten aufgerufenen Dokuments gespeichert. Es findet sich eine negative Korrelation zwischen dem durch die Probanden betriebenen Aufwand und der gemessenen Zufriedenheit. Die Teilnehmer sind also umso unzufriedener, je mehr Aufwand sie investieren müssen. Des Weiteren können Al-Maskari und Sanderson (ebd.) feststellen, dass die Zeit, die benötigt wird, um die gesuchte Information zu finden, einen direkten Einfluss auf die Benutzerzufriedenheit ausübt. Hier zeigt sich: Je kürzer die Suchdauer desto höher ist das abschließende Zufriedenheitsurteil. Zu diesem Schluss kommen auch Kelly et al. (2007) im Rahmen der ersten beiden Experimente ihrer Nutzerstudie (vgl. Abschn. 3.3.2.2). Beide Experimente unterscheiden sich ausschließlich in der Instruktion der Testpersonen. Während diese im Rahmen des ersten Experiments jeweils alle zehn Dokumente der Reihe nach bezüglich ihrer Relevanz für die entsprechende Aufgabe bewerten sollen, werden die Teilnehmer des zweiten Experiments gebeten, fünf relevante Dokumente zu finden. Sobald die Teilnehmer das fünfte Dokument als relevant markieren, werden sie zur Suchmaschinenevaluierungsseite weitergeleitet, sodass die Bearbeitungsdauer für die einzelnen Aufgaben unterschiedlich ausfallen kann. Indem sie die Zufriedenheitswerte des ersten mit denen des zweiten Experiments vergleichen, finden auch Kelly et al. (ebd.), wie in der Studie von Al-Maskari und Sanderson (2010), dass Probanden, die im zweiten Experiment kürzer suchen, höhere Zufriedenheitswerte aufweisen als Probanden, die im ersten Experiment alle zehn Dokumente bewerten müssen. Auch Xu und Mease (2009), die Bearbeitungszeiten von Benutzern für zwei Suchsysteme vergleichen, kommen zu dem Schluss, dass die Dauer einer Suche negativ mit der Suchzufriedenheit korreliert. Darüber hinaus können sie zeigen, dass die Bearbeitungszeit als Prädiktor gut zwischen Systemen unterschiedlicher Retrievalqualität differenziert. Die Untersuchungsmethodik stellt sich hierbei wie folgt dar: 200 Probanden werden zufällig in zwei Gruppen eingeteilt, die jeweils einem der beiden zu testenden Suchalgorithmen zugewiesen werden (mixed-design). Der Systemunterschied ist über ein nDCG-Maß realisiert, der jedoch nicht genauer präzisiert wird. Um den Einfluss der Bearbeitungszeit auf die Benutzerzufriedenheit zu berücksichtigen, bewerten die Probanden ihre Zufriedenheit auf einer Skala von 1 (very dissatisfied) bis 5 (very satisfied). Als weitere Verbesserung des Studiendesigns schlagen Xu und Mease (ebd.) vor, jeden Teilnehmer beide Systeme verwenden zu lassen, um den Einfluss der individuellen Suchleistung zu reduzieren.

Eine Studie von Kiseleva et al. (2016), die verschiedene Aspekte von Benutzerzufriedenheit im Kontext der mobilen Internetnutzung analysiert, kommt ebenfalls zu dem Ergebnis, dass der Aufwand, den ein Nutzer betreibt, in engem Zusammenhang mit der Benutzerzufriedenheit steht. Die Autoren führen drei Experimente durch, die neben dem Einfluss des zur Aufgabenerledigung benötigten Aufwands auf die Benutzerzufriedenheit auch die Komplexität der Testaufgaben sowie die Qualität der Spracherkennungssoftware (Cortana) untersuchen. Jedem der Experimente liegt dabei ein anderes Nutzungszenario (Gerätebedienung, Websuche u. Suchdialog) zugrunde. Nach einer Trainingsaufgabe bearbeiten die Teilnehmer jeweils eine Reihe von Testaufgaben, die sie selbstbestimmt beenden können. Im Anschluss an jede Aufgabe bewerten sie ihren Sucherfolg, ihre Zufriedenheit, den wahrgenommenen Aufwand sowie die Qualität der Spracherkennung. In allen drei Experimenten kann eine negative Korrelation zwischen Suchzufriedenheit und wahrgenommenem Aufwand nachgewiesen werden.

Nachdem nun mehrere Studien betrachtet wurden, die einen Zusammenhang zwischen der aufgewendeten Suchzeit und der resultierenden Suchzufriedenheit nachweisen, wird im Folgenden eine aktuelle Studie vorgestellt, die auf diese Frage interessante Antworten gibt. Diese Studie von Luo et al. (2017) basiert auf einem kontrollierten Laborexperiment, in dem die Autoren den Effekt der Bearbeitungszeit auf die von den Testpersonen wahrgenommene Bearbeitungsdauer untersuchen. An dieser Studie nehmen 50 Probanden teil. Sie bearbeiten 9 adhoc Suchaufgaben mit unterschiedlichen Schwierigkeitsgraden, wobei die erste Aufgabe als Trainingsaufgabe fungiert. Um den Einfluss der Suchergebnisse zu minimieren, verwenden die Teilnehmer vordefinierte Suchanfragen und erhalten in Bezug auf Precision und Ranking manipulierte Suchergebnisse. Alle Probanden bearbeiten vier Aufgaben mit hoher und vier Aufgaben mit niedriger Systemqualität (Within-Subject-Design). Neben ihrer Zufriedenheit mit den Suchergebnissen werden die Testpersonen gebeten, ein retrospektives Zeiturteil bezüglich der Bearbeitungsdauer abzugeben. Die Ergebnisse legen nahe, dass Teilnehmer dazu tendieren lange Aufgabenzeiten kürzer wahrzunehmen, während sie kurze Aufgabenzeiten häufig überschätzen. Nach Aufteilung der Probanden in zufriedene und unzufriedene Teilnehmer, können Luo et al. (ebd.) interessanterweise beobachten, dass dieser Wahrnehmungseffekt in der ersten Gruppe stärker ausgeprägt zu sein scheint. Darüber hinaus können die Autoren zeigen, dass die Aufgabenschwierigkeit einen signifikanten Einfluss auf das Zeiturteil hat, der dazu beiträgt, dass die Probanden bei schwierigeren Aufgaben dazu neigen, die Bearbeitungsdauer länger einzuschätzen als sie tatsächlich ist. Luo et al. (ebd., S. 134) führen diese Beobachtungen auf ein häufig in der Motivationspsychologie auftretendes Phänomen zurück, dass positive Erfahrungen zu einer Unterschätzung des Zeitaufwands führen, während negative Erfahrungen eine Überschätzung zur Folge haben. Insgesamt könnten diese Befunde darauf hindeuten, dass zufriedene Nutzer den Aufwand ihrer Suche im Nachhinein geringer einschätzen.

Eine letzte Studie, die in diesem Zusammenhang betrachtet werden soll, ist die Arbeit von Sandore, 1990. In dieser als Telefonumfrage durchgeführten Erhebung mit 171 Nutzern einer öffentlichen Bibliothek stellt die Autorin fest, dass Nutzer, die im Zuge einer Auftragsrecherche nicht anwesend sind, im Durchschnitt zufriedener sind als Nutzer, die während einer solchen Recherche persönlich zugegen sind. Die Autorin vermutet, dass der Grund hierfür in einem zu geringen Verständnis des Suchprozesses liegen könnte: „They most often expressed appreciation

for efficient service, the mechanics of which they were not required to understand. Perhaps these users were most appreciative because they were most removed from the search process, and tended to view it as a somewhat mysterious, black-box procedure. “ (Sandore, 1990, S. 45) Wie die zuvor beschriebenen Studien zeigen konnten, spielen auch die benötigte Zeit und der entsprechende Aufwand bei der Qualitätswahrnehmung eine wichtige Rolle. Eine weitere Erklärung könnte also auch sein, dass Teilnehmer, die während des Rechercheprozesses nicht anwesend sind, deshalb zufriedener sind, weil sie praktisch keinen Aufwand mit der Suche haben.

Die im Folgenden vorgestellten Studien nehmen die durch die Nutzer erreichte Suchleistung in den Blick. Neben dem Einfluss von Systemleistung und Benutzeraufwand auf die Wahrnehmung von Suchergebnissen gehen Al-Maskari und Sanderson, 2010 auch der Frage nach, in welcher Weise die individuell erbrachte Benutzerleistung für die Zufriedenheitseinschätzung bedeutsam ist. Dazu erfassen sie sowohl die Anzahl der von den Testpersonen in Übereinstimmung mit den TREC-Juroren als relevant eingestuften Dokumente als auch die insgesamt von den Testpersonen als relevant bewerteten Dokumente. Al-Maskari und Sanderson (ebd.) testen auch den Einfluss dieser Variablen mithilfe eines Korrelationstests. Im Ergebnis zeigt der Spearman-Korrelationskoeffizient, dass für beide Benutzerleistungsmaße ein positiver Zusammenhang zwischen der erbrachten Leistung und der subjektiv empfundenen Zufriedenheit besteht.

In einer früheren Studie finden Al-Maskari et al. (2006) keinen direkten Zusammenhang zwischen Benutzerleistung und Benutzerzufriedenheit. Diese Studie wird im Rahmen einer Teilnahme am interactive track von CLEF durchgeführt. Im Jahr 2006 ist in diesem track das zentrale Thema Bildretrieval und Fotos der Fotocommunity Flickr dienen als Dokumentenkollektion. Die Systemleistung wird in dieser Studie über fünf Effektivitätsmaße erfasst ($P@50$norm; $P@100$norm; Q-measure; BPref-10; 10-Precision). Die individuelle Suchleistung von 11 Testpersonen wird mithilfe der auch in dieser Arbeit verwendeten Benutzermaße für Recall und Precision erhoben (vgl. Abschn. 4.2.2.2). Darüber hinaus befragen Al-Maskari et al. (ebd.) die Teilnehmer in Bezug auf ihre Zufriedenheit mit der Nützlichkeit, der Genauigkeit und der Vollständigkeit der Suchergebnisse, die diese anhand einer dreistufigen Skala bewerten. Es zeigt sich, dass weder zwischen der zugrunde liegenden Systemqualität und der individuellen Suchleistung noch zwischen der Suchleistung und der Benutzerzufriedenheit nennenswerte Korrelationen berichtet werden können. Diese mangelnde Übereinstimmung zwischen den erhobenen Benutzermaßen und der Systemgüte führen Al-Maskari et al. (ebd.) auf die Eignung precision-orientierter Systemleistungsmaße zurück: „The lack of correlation between users and the system as determined by precision-oriented metrics indicate that these metrics are not compatible with user satisfaction and performance.“ (ebd., S. 3) Allerdings sollte hier kritisch angemerkt werden, dass sowohl die geringe Zahl der Testpersonen als auch die ungewöhnliche Bildretrievalaufgabe die Generalisierbarkeit der Ergebnisse auf den allgemeinen Suchkontext einschränken.

Abschließend ist zu bemerken, dass auch die Wahrnehmung der eigenen Suchleistung einen Einfluss auf die Zufriedenheit mit einem Suchsystem ausüben kann. Dies betrifft zum einen die in Abschnitt 3.2.2 vorgestellte Studie von Jiang und Allan (2016). Sie können nachweisen, dass der schlechteste nDCG-Werte im Verlauf einer Suchsession am stärksten mit der Einschätzung der eigenen Suchleistung korreliert. Dies lässt Jiang und Allan (ebd., S. 288) in Bezug auf die wahrgenommene Systemqualität vermuten, dass „[. . .] a few underperforming queries in a session

may substantially affect user experience, while it is common to find well-performing queries in any session.“ Su (1994) bringt einen weiteren Aspekt des wahrgenommenen Sucherfolgs in die Diskussion ein (vgl. Abschn. 3.3.2.1 u. 3.3.2.2). Neben klassischen Evaluierungsmaßen, die sich den Bereichen Relevanz, Effizienz und Nützlichkeit zuordnen lassen, erhebt Su (ebd.) u.a. das Vertrauen bzw. die Zuversicht der Benutzer in die Vollständigkeit der erreichten Suchergebnisse. Von den insgesamt 20 in dieser Studie abgefragten Evaluierungsitems korreliert das Vertrauen der Testpersonen in die Vollständigkeit der Suche am drittstärksten mit dem insgesamt wahrgenommenen Sucherfolg. Dies kann als Indiz gewertet werden, dass auch der empfundene Recall Einfluss auf die Nutzerzufriedenheit hat.

Die hier vorgestellten Studien lassen insgesamt den Schluss zu, dass sowohl der Nutzeraufwand als auch die Nutzerleistung mit der Nutzerzufriedenheit korrelieren. Insbesondere für die Bearbeitungszeit kann dieser Zusammenhang in mehreren unabhängigen Studien nachgewiesen werden. Diese Tatsache ist auch von praktischer Relevanz, da sich Bearbeitungszeiten im Gegensatz zu anderen Benutzermaßen auch in realen Suchkontexten mit Hilfe von Logdateien erheben lassen. Für den Zusammenhang zwischen Sucherfolg und Nutzerzufriedenheit ergibt sich hingegen ein weniger einheitliches Bild und der Effekt selber kann als weniger gesichert gelten. Ein Hauptziel der vorliegenden Arbeit ist es deshalb, diesen Zusammenhang weiter zu erforschen und aufbauend auf einer größeren Gruppe von Benutzerleistungsmaßen mögliche Korrelationen zu analysieren. Aus Sicht der nutzerzentrierten IR-Evaluierung scheint es jedoch geboten, auch kontextuelle Ursachen für das Zufriedenheitsempfinden mit in die Betrachtung einzubeziehen. Aus diesem Grund stellt der folgende Abschnitt einige Studien vor, die eine prozessorientierte Sichtweise auf die Entstehung von Nutzerzufriedenheit verfolgen.

3.3.2.4. Zufriedenheit aus dynamischer Perspektive

Dieser Abschnitt widmet sich der Frage nach der Kontextabhängigkeit und Dynamik von Zufriedenheitsurteilen sowie der daraus folgenden Relativität von Qualitätszuschreibungen. Mit der Hinwendung zu benutzerorientierten Evaluierungsansätzen hat das IR eine neue Ausrichtung erfahren. Im Gegensatz zu dem klassischen, systemorientierten Paradigma handelt es sich beim IIR um einen dynamischen, prozessorientierten Ansatz, der über eine rein objektive Bestimmung der Systemqualität hinaus geht. Es stellt die einzelnen Elemente des IR in einen zeitlichen Kontext, fügt mit dem Benutzer und dessen Interaktion mit dem System aber noch weitere Aspekte hinzu, die wiederum das zugrunde gelegte Konzept der Relevanz bedingen (vgl. Abschn. 3.1.1). Wie in Abschnitt 3.3.1.4 bereits dargestellt, ist angesichts dieses Prozesscharakters auch bei der Qualitätswahrnehmung der Benutzer davon auszugehen, dass sich das Zufriedenheitsurteil aus mehreren, phasenbezogenen Teilzufriedenheiten zusammensetzt. Im Laufe der Interaktion können so z.B. andere Aspekte des Suchthemas in den Vordergrund des Interesses rücken. Die Benutzer können das Thema komplexer wahrnehmen oder dieses durch die Einordnung in übergeordnete Kategorien besser verstehen (vgl. Abschn. 3.1.3). Dieser Abschnitt konzentriert sich in diesem Zusammenhang auf die Frage, wie die Weiterentwicklung des Benutzers im Rahmen interaktiver IR-Experimente berücksichtigt werden kann und welchen Einfluss dies auf die Gesamtbeurteilung der Suche hat.

Die Frage, wie kontinuierliche Reflexionsprozesse im Zeitverlauf das Zufriedenheitsempfinden beeinflussen, wird zwar in einer Reihe von Studien zur Benutzerzufriedenheit implizit oder

explizit angesprochen (Smithson, 1994; Bruce, 1994; White et al., 2010; Hu et al., 2011), jedoch bisher nur selten systematisch untersucht. Zwei aktuelle Studien setzten sich jedoch mit Teilaspekten dieser Thematik auseinander und sollen im Folgenden vorgestellt werden. Untersuchungsgegenstand ist dabei einerseits der dynamische Zusammenhang zwischen wahrgenommener Ergebnisqualität, individuellem Aufwand und resultierender Gesamtzufriedenheit (Jiang et al., 2015) und andererseits der in der Einleitung angesprochene dynamische Charakter der Relevanzwahrnehmung (Jiang et al., 2017). Beide Studien sind vor allem auch aus methodischer Sicht relevant, da sehr unterschiedliche Herangehensweisen zur Erhebung der Qualitätswahrnehmung im Zeitverlauf herangezogen werden.

Im Fall der ersten Studie handelt es sich um eine Logfile-Analyse von 476 Suchsitzungen, die im Jahr 2014 mit der Suchmaschine Bing durchgeführt werden. Jiang et al. (2015) engagieren zu diesem Zweck crowd worker, die die einzelnen Suchsitzungen stellvertretend für die ursprünglichen Benutzer bewerten. Diese Annotatoren bewerten sowohl die Suchzufriedenheit in Bezug auf die gesamte Suchsitzung auf einer 5-stufigen Skala (*very unsatisfied* bis *very satisfied*) als auch die Qualität der Suchergebnisseiten in Bezug auf die einzelnen Suchanfragen auf einer 3-stufigen Skala (*very*, *somewhat* u. *not useful*). Pro Suchsitzung werden drei Annotationen erfasst, die anschließend gemittelt werden. Darüber hinaus wird für jede Suchsitzung der Aufwand anhand der Anzahl der Suchanfragen pro Session ermittelt. Hinsichtlich einer phasenbezogenen Zuordnung der Suchzufriedenheiten ermitteln Jiang et al. (ebd.) Unterschiede in Bezug auf die wahrgenommene Ergebnisqualität und den individuellen Aufwand zwischen der ersten und letzten Suchanfrage einer Suchsitzung und korrelieren diese mit den entsprechenden Gesamtzufriedenheiten. Zunächst zeigt sich, in Übereinstimmung mit Abschnitt 3.3.2.3, dass der mit dem Suchaufwand gewichtete Sucherfolg die stärkste Korrelation mit der Gesamtzufriedenheit aufweist. In Bezug auf die dynamische Entwicklung geht hervor, dass eine Zunahme der wahrgenommenen Ergebnislistenqualität zwischen der ersten und letzten Suche positiv mit dem Zufriedenheitsurteil korreliert. Dies lässt sich dahingehend interpretieren, dass die Gesamtzufriedenheit mit dem Sucherlebnis maßgeblich von der letzten durchgeführten Suche beeinflusst zu werden scheint. In diesem Sinne lassen sich die von Jiang et al. (ebd.) auf Ebene der Suchanfragen erfassten Ergebnisqualitäten als phasenbezogene Teilzufriedenheiten begreifen, die auf Ebene der Suchsitzungen in die Gesamtzufriedenheit einfließen.

Die zweite Studie beruht auf einem klassischen IIR-Experiment mit 28 Probanden und 28 Suchaufgaben des TREC session track. Davon bearbeitet jeder Teilnehmer 4 Aufgaben, sodass pro Aufgabe die Interaktionsdaten von 4 Teilnehmern vorhanden sind. Insgesamt ergeben sich somit 112 Suchsitzungen und 736 von den Probanden angesehene Dokumente. Um die dynamische Weiterentwicklung der Testpersonen untersuchen zu können, bewerten die Probanden jedes angesehene Dokument hinsichtlich der Kriterien *Neuheitswert*, *Aufwand*, *Verständlichkeit* und *Vertrauenswürdigkeit* sowie in Bezug auf die situativ wahrgenommene Relevanz (*How much useful information did you get from this webpage?*) (Jiang et al., 2017, S. 140). Diese situationsbedingte Definition von Relevanz bezeichnen die Autoren als *ephemeral state of relevance (ESR)*, die an die situative Relevanz von Saracevic angelehnt ist (Saracevic (1996), vgl. Abschn. 3.1.1). Im Anschluss an jede Suchsitzung werden die Teilnehmer nochmals über ihre Erfahrungen und Wahrnehmungen bei der Benutzung des Suchsystems befragt. Zum einen werden sie gebeten,

die zuvor angesehenen Dokumente erneut zu bewerten. Dabei umfassen die zur Neubewertung herangezogenen Kriterien neben Neuheitswert, Verständlichkeit und Vertrauenswürdigkeit diesmal auch die Kategorien thematische Relevanz und Nützlichkeit. Darüber hinaus werden sechs Sucherlebnismaße erfasst: Zufriedenheit, Frustration, Nützlichkeit des Systems, Zielerreichung, Aufwand und Aufgabenschwierigkeit. Auf Basis dieser Benutzerdaten analysieren Jiang et al. (2017, S. 144 f.) den Zusammenhang zwischen ESR, Nützlichkeitkeitsbewertungen und thematischen Relevanzurteilen auf der einen und den sechs Sucherlebnismaßen auf der anderen Seite. Die Ergebnisse zeigen, dass Nützlichkeitsbewertungen im Vergleich zu thematischen Relevanzurteilen tatsächlich stärker mit den sechs Sucherlebnismaßen korrelieren. Die während der Suche getroffenen ESR-Bewertungen zeigen jedoch bis auf die Frustration die stärkste Korrelation mit dem Sucherlebnis. Begreift man die einzelnen ESR-Bewertungen als Teilzufriedenheitsurteile im Sinne des dynamisierten C/D-Paradigmas (vgl. Abschn. 3.3.1.4), so kann dieses Ergebnis dahingehend interpretiert werden, dass die mittlere Teilzufriedenheit tatsächlich einen starken Einfluss auf die Bewertung des Gesamtsucherlebnisses hat. Allerdings fällt der Unterschied der Korrelationsstärke zwischen Nützlichkeits- und ESR-Bewertung eher gering aus, weswegen Jiang et al. (ebd.) bezweifeln, dass sich der zusätzliche Aufwand für die Testpersonen lohnt. Es ist jedoch anzumerken, dass nicht klar wird, ob die Abfrage des Sucherlebnisses vor oder nach der Neubewertungsphase stattfindet. Eine Abfrage des Sucherlebens im Anschluss an die Neubewertung könnte hier zu einem verzerrten Urteil geführt haben, da die tatsächliche Sucherfahrung durch die erneute Bewertung der Dokumente überdeckt worden sein könnte. Des Weiteren wäre es interessant gewesen, nicht nur die über die gesamte Session gemittelten ESR-Bewertungen heranzuziehen, sondern, ähnlich wie bei der zuvor beschriebenen Studie von Jiang et al. (2015), auch die Korrelation der ersten Bewertungen im Vergleich zu den letzten Bewertungen zu betrachten.

Abschließend soll noch einmal kurz auf eine bereits in Abschnitt 3.3.2.1 vorgestellte Studie von Szajna und Scamell (1993) eingegangen werden, die dynamische Aspekte der Benutzerzufriedenheit anspricht. Insbesondere können Szajna und Scamell (ebd.) nachweisen, dass unrealistische Erwartungen der Testteilnehmer bereits nach der ersten Interaktion mit einem Informationssystem angepasst werden und in den nachfolgenden Interaktionen stabil bleiben. In Bezug auf den Soll/Ist-Vergleich für die Teilzufriedenheiten im dynamischen C/D-Paradigma (vgl. Abschn. 3.3.1.4) ist somit ein besonders starker Effekt bzw. Unterschied für die ersten beiden Interaktionen zu erwarten. Allgemein bestätigt dieses Ergebnis auch Beobachtungen von Olson und Dover (1979), die zeigen können, dass Erwartungen abgebaut werden, wenn Testpersonen wiederholt Diskonfirmationen in Bezug auf diese Erwartungen ausgesetzt sind.

In der Gesamtschau zeigen die in diesem Abschnitt vorgestellten Studien, dass Benutzerzufriedenheit im Suchprozess tatsächlich eine dynamische Komponente aufweist und diese auch im Kontext von Nutzerstudien sichtbar gemacht werden kann. Dies wird zum einen deutlich am unterschiedlichen Einfluss der Suchergebnisqualität der ersten und letzten Suche auf die Gesamtzufriedenheit in der Studie von Jiang et al. (2015). Zum anderen kann auch die höhere Korrelationsstärke des ephemeral state of relevance (ESR) mit dem finalen Zufriedenheitsurteil in diesem Sinne gedeutet werden. Einen möglichen Ansatzpunkt zur Erklärung einer solchen Dynamik anhand des dynamisierten C/D-Paradigmas gibt die von Szajna und Scamell (1993)

beobachtete Änderung der Erwartungshaltung. Allerdings wird auch deutlich, dass weiterhin ein großer Forschungsbedarf besteht, um die genauen dynamischen Wirkungszusammenhänge im Kontext der Informationssuche zu verstehen. Die im Rahmen des empirischen Teils dieser Arbeit durchgeführten Nutzerstudien zielen darauf ab, hier weitere Aufklärung zu leisten.

3.4. Fazit: Bewertung von Suchergebnissen

Zusammenfassend wird in diesem Kapitel auf Grundlage theoretischer Konzepte und der aktuellen Studienlage eine prozess- und ergebnisorientierte Sichtweise auf die Bewertung von Suchergebnissen entwickelt. Dieses Vorgehen erscheint gerade vor dem Hintergrund der vielschichtigen Ergebnisse in Bezug auf die Übertragbarkeit von systemorientierten Ergebnissen auf den Nutzerkontext notwendig und sinnvoll. Des Weiteren trägt dieser Ansatz der Tatsache Rechnung, dass Suchergebnisse erst im Rahmen der Interaktion zwischen Nutzer und System zustande kommen und auf diese Weise nicht nur das Endergebnis, sondern auch sein Entstehungsprozess wahrgenommen und bewertet wird. Innerhalb dieser interaktiven Betrachtung des Retrievalprozesses ist es dem Benutzer möglich, direkt Einfluss zu nehmen – er ist nicht ausschließlich passiver Empfänger sondern aktiv, gemäß seiner individuellen Voraussetzungen, an diesem Prozess beteiligt. Dies hat zu Folge, dass nicht nur die Bewertung sondern auch der Erhalt von Suchergebnissen sowohl inter- als auch intraindividuell variiert. Eine Sichtweise, die durch die diskutierten Studien zur Dynamik des Suchprozesses gestützt wird.

Für die Interpretation interaktiver Retrievaltests ist es deshalb wichtig auch personen- und situationsbezogene Einflussfaktoren zu berücksichtigen. Eine besondere Rolle kommt hierbei der Relevanzwahrnehmung und ihrer Kontextabhängigkeit im Suchprozess zu. Um ein möglichst umfassendes Spektrum der Suchergebniswahrnehmung einzufangen und dem individuellen Charakter des interaktiven Suchprozesses gerecht zu werden, ist es darüber hinaus erforderlich, sowohl subjektive als auch objektive Maße zur Bestimmung des Sucherfolgs einzubeziehen. In Bezug auf die Nutzerzufriedenheit kann vor dem Hintergrund der theoretischen Erklärungsmodelle aus der Kundenzufriedenheitsforschung in Verbindung mit den betrachteten Studien zur Informationssuche davon ausgegangen werden, dass insbesondere auch die Erwartungshaltung der Nutzer einen Erklärungsbeitrag liefert. Darüber hinaus erlaubt die Studienlage weitere Determinanten von Nutzerzufriedenheit zu identifizieren. Diese umfassen neben individuellen Faktoren wie dem Nutzeraufwand und dem wahrgenommenen Sucherfolg auch die objektiv gemessene Systemqualität. Die zentralen Erkenntnisse dieses Kapitels sind noch einmal in Tabelle 3.1. Aufbauend auf diesen Ergebnissen wird im nun folgenden Kapitel 4 das methodische Vorgehen zur Untersuchung der in Abschnitt 1.2 dargestellten Forschungsfragen entwickelt. Neben der praktischen Umsetzung der Testgestaltung geht es dabei insbesondere um eine zielführende Operationalisierung der experimentellen Parameter und Messgrößen.

Tab. 3.1.: Zentrale Schlussfolgerungen zur Bewertung von Suchergebnissen.

Relevante Erkenntnisse	Autoren	Implikationen
Kriterien der Relevanzbeurteilung		
Ansätze zur Identifikation allgemeiner Kriterien der Relevanzbeurteilung kommen zu ähnlichen Ergebnissen	Schamber (1991), Park (1993), Barry (1994), Schamber (1994), Fidel und Crandall (1997), Barry und Schamber (1998), Fitzgerald und Galloway (2001) und Maglaughlin und Sonnenwald (2002)	Existenz eines begrenzten Kriterienkanons
Anwendung von Relevanzkriterien situativ und personenspezifisch	Howard (1994)	Der Benutzer als zentraler Einflussfaktor im IR-Prozess
Nur geringe Überschneidungen beim Suchen und Auffinden relevanter Dokumente durch unterschiedliche Suchende	Saracevic et al. (1988), Saracevic und Kantor (1988a) und Saracevic und Kantor (1988b)	Vielschichtigkeit des Konstrukts der Relevanz
Benutzer verwenden Entscheidungsregeln, die dabei helfen, relevante Dokumente auszuwählen	Wang und Soergel (1998) und Wang und White (1999)	Entscheidungsregeln liefern Handlungssystematik, um Auswahlverhalten von Nutzern klassifizieren zu können
Dynamische Aspekte bei der Relevanzbeurteilung		
Untersuchungen zeigen heterogene Ergebnisse in Bezug auf dynamische Aspekte bei der Relevanzbeurteilung	Bruce (1994), Smithson (1994), Tang und Solomon (1998), Wang und Soergel (1998), Wang und White (1999), Vakkari und Hakala (2000), Vakkari (2001), Tang und Solomon (2001) und Smucker und Jethani (2010a)	Ergebnisse müssen unter besonderer Berücksichtigung des Studiendesigns bewertet werden
Schwelle, ab wann ein Suchergebnis als relevant eingestuft wird, bestimmt sich nicht nur aus den übergeordneten Relevanzkriterien, sondern ist immer auch personenabhängig	Wang und Soergel (1998), Scholer und Turpin (2008) und Scholer et al. (2008)	Individuelle Relevanzschwellenwerte als Ursache für Unterschiede zwischen system- und benutzerorienten Untersuchungen
Aktuelle Verfügbarkeit und Phase im Suchprozess als Gründe für Anpassung relevanzbezogener Maßstäbe	Wang und Soergel (1998), Wang und White (1999) und Smucker und Jethani (2010a)	Rolle des situativen Kontexts bei der Festlegung individueller Relevanzschwellen
Vertrauen in die eigenen Relevanzurteile nimmt im Suchverlauf zu	Tang und Solomon (1998) und Tang und Solomon (2001)	Lerneffekte können zu einem veränderten Bewertungsverhalten führen
Übertragbarkeit systemorientierter Ergebnisse		
Untersuchungen zeigen heterogene Ergebnisse im Zusammenhang mit der Übertragbarkeit systemorientierter Ergebnisse auf den Benutzerkontext	Turpin und Hersh (2001), Allan et al. (2005), Turpin und Scholer (2006), Al-Maskari et al. (2008b), Al-Maskari et al. (2008a) und Smucker und Jethani (2010a)	Ergebnisse müssen unter besonderer Berücksichtigung des Studiendesigns bewertet werden
Je höher der relative Systemunterschied, desto wahrscheinlicher eine positive Korrelation zwischen System- und Benutzerleistung	Allan et al. (2005), Al-Maskari et al. (2008b), Al-Maskari et al. (2008a) und Smucker und Jethani (2010a)	Betrachtung des relativen Systemunterschieds erforderlich
Benutzer sind insbesondere in Bezug auf recallorientierte Benutzerleistungsmaße in der Lage, systembedingte Qualitätsunterschiede zu kompensieren	Turpin und Hersh (2001) und Smith und Kantor (2008)	Prüfung der Übertragbarkeit systemorientierter Ergebnisse im Kontext unterschiedlicher Leistungsparameter geboten
Anwendung von Relevanzkriterien situativ und personenspezifisch	Turpin und Hersh (2001), Smith und Kantor (2008), Scholer und Turpin (2009) und Smucker und Jethani (2010a)	Fundierung der Kontextabhängigkeit der Relevanzwahrnehmung durch weitere kontrollierte Studien notwendig

Fortsetzung auf nächster Seite

Tab. 3.1 Zentrale Schlussfolgerungen zur Bewertung von Suchergebnissen (Fortsetzung)

Relevante Erkenntnisse	Autoren	Implikationen
Wahrnehmung des Sucherfolgs		
Benutzer sind in der Lage, Qualitätsunterschiede wahrzunehmen	Thomas und Hawking (2006), Kelly et al. (2007) und Al-Maskari und Sanderson (2010)	Wahrgenommener Sucherfolg als Determinante für Benutzerzufriedenheit relevant
Gefühl der Informationssättigung ab einem Recallniveau von etwa 60%	Turpin und Hersh (2001), Allan et al. (2005), Al-Maskari et al. (2006) und Dostert und Kelly (2009)	Informationssättigung als mögliche Erklärung für heterogene Ergebnisse im Zusammenhang mit der Übertragbarkeit systemorientierter Ergebnisse
Wahrnehmung des Sucherfolgs korreliert mit Suchleistung, Selbsteinschätzung weicht aber vom tatsächlichen Recall-Wert ab	Al-Maskari et al. (2006) und Dostert und Kelly (2009)	Alleinige Fokussierung auf Benutzerzufriedenheit und Selbsteinschätzung kann zu kurz greifen, deshalb auch direkte Erhebung von Benutzerleistung notwendig
Die wahrgenommene Systemqualität ist eine wichtige Einflussgröße für die Benutzerzufriedenheit	Cox und Fisher (2004), Kwahk und Oh (2009) und Al-Maskari und Sanderson (2010)	Änderungen in der Relevanzwahrnehmung sollten sich auch in der Benutzerzufriedenheit widerspiegeln
Wahrnehmung des Sucherfolgs abhängig von situativen Faktoren	Jiang und Allan (2016)	Bewertung von Suchergebnissen immer als Prozess analysieren
Einfluss von Erwartungen auf die Zufriedenheit		
Benutzererwartungen weisen einen Zusammenhang zur Benutzerzufriedenheit auf	Szajna und Scamell (1993), Su (1994), Cox und Fisher (2004), Jansen et al. (2007) und Kwahk und Oh (2009)	Macht Erwartungshaltung als Determinante für Benutzerzufriedenheit plausibel
Hohe Markenerwartungen tragen zu einer positiveren Qualitätswahrnehmung der Suchergebnisse bei	Jansen et al. (2007)	Möglicher Hinweis auf Nicht-Übertragbarkeit des C/D-Paradigmas auf Kontext der Informationssuche
Beobachtete Überschätzung der eigenen Suchleistung	Kelly et al. (2008b) und Kelly et al. (2008a)	Begrenzte Objektivität bei der Zufriedenheitsbeurteilung (Eigener Beitrag wird mitbewertet)
Vergleiche zwischen Systemqualität und Zufriedenheit		
Korrelation zwischen Systemleistung und Benutzerzufriedenheit	Johnson et al. (2003), Thomas und Hawking (2006), Huffman und Hochster (2007), Al-Maskari et al. (2007) und Kelly et al. (2007)	Unterschiede in der Systemqualität im Zufriedenheitsurteil der Benutzer nachweisbar
Ein positiver Zusammenhang zwischen Precision und Benutzerzufriedenheit kann in vielen Studien nachgewiesen werden	Su (1994), Su (2003), Johnson et al. (2003), Kelly et al. (2007) und Al-Maskari et al. (2007)	Bedeutung der Precision für die Benutzerzufriedenheit
Cumulative Gain Maße korrelieren positiv mit dem finalen Zufriedenheitsurteil der Benutzer	Huffman und Hochster (2007) und Al-Maskari et al. (2007)	Operationalisierung der Systemunterschiede sollte auch in Bezug auf Cumulative Gain Maße sichtbar sein
Zufriedenheitsurteil spiegelt sich im Suchverhalten der Nutzer wider	Thomas und Hawking (2006), Kelly et al. (2007) und Al-Maskari und Sanderson (2010)	Konsistenz von Zufriedenheitsurteilen und Nutzerverhalten kann überprüft werden

Fortsetzung auf nächster Seite

Tab. 3.1 Zentrale Schlussfolgerungen zur Bewertung von Suchergebnissen (Fortsetzung)

Relevante Erkenntnisse	Autoren	Implikationen
Zusammenhang zwischen Sucherfolg und Zufriedenheit		
Untersuchungen zeigen heterogene Ergebnisse in Bezug auf den Zusammenhang zwischen der erbrachten Benutzerleistung und der subjektiv empfundenen Benutzerzufriedenheit	Al-Maskari et al. (2006) und Al-Maskari und Sanderson (2010)	Bedarf einer Klärung durch weitere Studien
Negative Korrelation zwischen dem durch den Benutzer betriebenen Aufwand zur Befriedigung des Informationsbedürfnisses und der gemessenen Zufriedenheit	Kelly et al. (2007), Xu und Mease (2009), Al-Maskari und Sanderson (2010) und Kiseleva et al. (2016)	Suchaufwand als weitere Determinante der Benutzerzufriedenheit
Je kürzer die Suchdauer desto höher das abschließende Zufriedenheitsurteil	Kelly et al. (2007), Xu und Mease (2009) und Al-Maskari und Sanderson (2010)	Bearbeitungszeit als Prädiktor zur Vorhersage des Zufriedenheitsurteils
Die Wahrnehmung von Bearbeitungszeiten hängt von Zufriedenheit und Aufgabenschwierigkeit ab	Luo et al. (2017)	Gegenseitige Abhängigkeit von Sucherfolgswahrnehmung und Zufriedenheit
Zufriedenheit aus dynamischer Perspektive		
Heterogener Einfluss der Suchergebnisqualität auf die Gesamtzufriedenheit im Suchverlauf	Jiang et al. (2015) und Jiang et al. (2017)	Benutzerzufriedenheit weist dynamische Komponente auf, die in Nutzerstudien sichtbar gemacht werden kann
Unrealistische Erwartungen werden abgebaut, wenn Nutzer Diskonfirmation in Bezug auf Erwartungen erfahren	Olson und Dover (1979) und Szajna und Scamell (1993)	Möglicher Ansatzpunkt zur Erklärung des heterogenen Einflusses der Suchergebnisqualität im Suchverlauf

4. Methodisches Vorgehen

Das vorliegende Kapitel stellt den methodischen Unterbau der im Rahmen dieser Arbeit durchgeführten Benutzerstudien bereit. Dieser umfasst neben allgemeinen Ausführungen zum Design von Nutzerstudien im IR insbesondere Ansätze zur Messung von Relevanz, Sucherfolg und Nutzerzufriedenheit sowie einen kurzen Überblick über die zur Auswertung herangezogenen statistischen Verfahren. Ausgehend von den in den vorangegangenen Kapiteln vorgestellten Studien liegt das Hauptaugenmerk darauf, ein angemessenes methodisches Vorgehen für die in der Einleitung vorgestellten Forschungsfragen zu entwickeln. Die tatsächliche experimentelle Umsetzung wird hingegen in den Kapiteln 5 bis 7 vorgestellt.

Konkret wird zunächst in Abschnitt 4.1 die Methodik des Laborexperiments im Kontext von IR-Evaluierungen diskutiert. Neben einer Betrachtung der Vor- und Nachteile dieses Ansatzes liegt der Fokus auf den in der Forschungsliteratur dokumentierten Testdesigns sowie deren Eignung für das vorliegende Forschungsvorhaben. Bevor abschließend in Abschnitt 4.3 die zur Auswertung herangezogenen statistischen Verfahren vorgestellt werden, geht Abschnitt 4.2 genauer auf die Operationalisierung der für diese Arbeit bedeutsamen Untersuchungsvariablen ein. In diesem Zusammenhang werden neben der Manipulation von Erwartungshaltung und Systemgüte sowie der Erfassung von Nutzerleistung und -zufriedenheit auch praktische Fragen zur Kontrolle von Störvariablen diskutiert.

4.1. Das Laborexperiment als Forschungsdesign im IIR

Ziel des interaktiven IR-Paradigmas ist es, das Verhalten realer Nutzer im Suchprozess zu verstehen. Ähnlich wie in der Psychologie und den Sozialwissenschaften führt die Berücksichtigung der Benutzer jedoch dazu, dass es oft schwierig ist, den Einfluss einzelner Faktoren in der realen Nutzungssituation zu isolieren und zu analysieren. Aus diesem Grund bieten sich auch im Kontext der IR-Evaluierung Studien mit Hilfe von Laborexperimenten an, die unter gut kontrollierten Bedingungen den realen Suchprozess so weit wie möglich nachbilden und gleichzeitig eine Fokussierung auf die für die untersuchte Forschungsfrage relevanten Aspekte erlauben.

In gewisser Weise lässt sich das Laborexperiment als Forschungsdesign im IIR auch als natürliche Weiterentwicklung des traditionellen systemorientierten Evaluierungsansatzes verstehen. Während zuvor Evaluierungsinitiativen standardisierte Korpora und Testaufgaben bereitstellen, um verschiedene Systeme direkt miteinander vergleichbar zu machen, wird in experimentellen Studien zum Informationssuchverhalten durch die Vorgabe von Suchaufgaben ein Rahmen geschaffen, in dem die Suchleistung und das Suchverhalten verschiedener Testpersonen miteinander verglichen werden können. Des Weiteren bieten Laborexperimente die Möglichkeit, zusätzliche Hintergrundinformationen, wie Vorerfahrungen und die Zufriedenheit mit den Suchergebnissen zu erfassen, die im Rahmen anderer Untersuchungstypen wie etwa Logfileanalysen

nicht ohne Weiteres zugänglich sind. Die Untersuchungsziele experimenteller Studien zum Informationssuchverhalten sind vielfältig und reichen von einfachen statistischen Analysen des Zusammenhangs von System- und Benutzerleistung über relevanzanalytische Fragestellungen bis hin zur Isolierung einzelner Einflussfaktoren und deren Bedeutung für das Informationssuchverhalten.

Da das Laborexperiment im Rahmen der vorliegenden Arbeit das Mittel der Wahl darstellt, wird im Folgenden ein kurzer Überblick über etablierte methodische Ansätze für Laborstudien und Logdatenanalysen im IIR-Kontext gegeben, sowie allgemeine Vor- und Nachteile der Methode diskutiert. Im Anschluss an eine Beschreibung der beiden Grundtypen experimenteller Forschungsdesigns folgt dann eine Darstellung der einzelnen Planungsschritte zur Durchführung eines Laborexperiments im IIR-Kontext.

4.1.1. Methodische Ansätze zur Analyse interaktiver Retrievalprozesse

Nachdem in den vorangegangenen Kapiteln Nutzerstudien in Bezug auf ihre Forschungsfragen und Ergebnisse analysiert werden, legt der nun folgende Abschnitt den Fokus auf ihre methodischen Herangehensweisen. Dabei ist das Ziel, ein geeignetes Vorgehen für die in Abschnitt 1.2 dargestellten Forschungsfragen zu identifizieren. Die wesentlichen Designentscheidungen die zur Planung einer Laborstudie getroffen werden müssen, umfassen insbesondere die Art der Suchaufgaben, das gewählte Testsystem und die Datenerhebung. Im Folgenden werden diese Aspekte in Bezug auf die Kriterien Realitätsgrad der Suchsituation, Kontrolle der Einflussparameter und zugängliche Interaktionsdaten diskutiert.

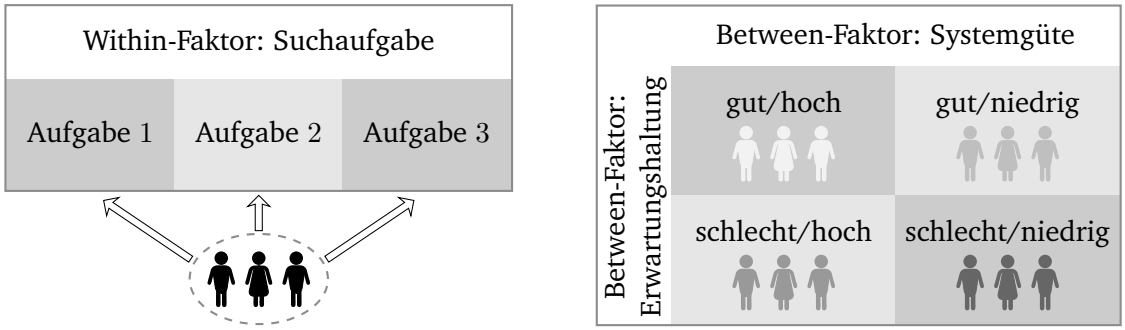
Ähnlich wie in der Psychologie und den Sozialwissenschaften können IIR-Studien zunächst anhand des gewählten experimentellen Typus unterteilt werden. Dabei lassen sich in Analogie zu Feld- und Laborexperimenten Studien, die auf Logdaten realer Suchsessions beruhen, von geplanten Nutzerexperimenten unterscheiden. Zwar weisen erstere in natürlicher Weise einen hohen Realitätsgrad sowie eine hohe externe Validität auf, erlauben auf der anderen Seite jedoch nur ein beschränktes Maß an Kontrolle und zugänglichen Nutzerdaten. So lassen sich typischerweise die verwendeten Suchbegriffe sowie die angesehen Informationsobjekte und entsprechenden Verweildauern bestimmen, allerdings bleibt offen, in welchem Maße das Informationsbedürfnis der Nutzer tatsächlich erfüllt wird bzw. welches initiale Informationsbedürfnis der Suchsession zugrunde liegt. Dies erschwert es einerseits, die Relevanz der betrachten Dokumente zu bewerten, während andererseits auch die subjektive Wahrnehmung der Sucherfahrung nicht erhoben werden kann. Einen Vorteil stellt hingegen die oft breite Datenbasis der Studien im Vergleich zu Laborexperimenten dar. Beispielhaft ist hier die Studie von White und Drucker (2007) zu nennen, die über eine Browsererweiterung über einen Zeitraum von fünf Monaten das Navigationsverhalten von gut 2500 Internetnutzern verfolgen, um typische Interaktionsstrategien zu identifizieren. Häufig kommt diese Methode auch zum Einsatz, um das Nutzerverhalten in Bezug auf ein real existierenden Suchsystem zu untersuchen (Jansen et al., 2000; Rieh u. Xie, 2001; Wang et al., 2003; Beitzel et al., 2004; Park et al., 2005; Jansen u. Spink, 2006). In diesem Zusammenhang stellen Logdatenanalysen in gewisser Weise einen Ex-Post-Fakto-Ansatz dar, da sie auf Grundlage der bereits vorhanden Interaktionsdaten im Nachhinein das Nutzerverhalten klassifizieren. Dies geht im Umkehrschluss mit einer fehlenden Kontrolle der genauen Untersuchungsbedingungen einher, die dieses Vorgehen zusammen mit der fehlenden individuellen Rückmeldung der Test-

personen für die in dieser Arbeit betrachteten Fragestellungen nicht sinnvoll erscheinen lassen. Insbesondere erscheint es in solch einem offen Kontext extrem schwierig bzw. von einem praktischen Gesichtspunkt her unmöglich, die Benutzerleistung der Testteilnehmer zu bestimmen.

Nutzerstudien hingegen, die den Suchprozess unter Laborbedingungen direkt beobachten, erlauben hier sowohl ein höheres Maß an Kontrolle, als auch den Zugriff auf umfassendere Nutzerdaten, was insbesondere Befragungen in Bezug auf das persönliche Erleben und die Einschätzung der eigenen Suchleistung mit einschließt. Im Prinzip ließe sich die Datenbasis in einem Laborexperiment noch durch weitere Parameter anreichern, die über die reine Befragung der Teilnehmer hinausgehen. Hier böte sich bspw. die aus der Usability-Forschung bekannte Methode des Eyetracking an, die weiteren Aufschluss über die Wahrnehmung von Ergebnislisten und das Leseverhalten aufgerufener Dokumente liefern kann: „Eye tracking adds meaning to the more traditional log file or click behavior analysis." (Pan et al., 2007). Zu nennen sind hier die Studien von Granka et al. (2004), Pan et al. (2004), Tullis (2007), Zaphiris und Savtich (2008) sowie Hill et al. (2011), die das Fixationsverhalten der Probanden beim Lesen von Ergebnislisten und aufgerufenen Dokumenten untersuchen. Kelly (2009, S. 199 f.) merkt jedoch an, dass diese Methode zwar große Mengen an Daten generiert, ihre Analyse und Interpretation sich jedoch häufig als schwierig herausstellt. Des Weiteren birgt das zusätzliche technische Equipment, welches zur Erhebung der Daten erforderlich ist, die Gefahr, die Testsituation in den Vordergrund treten zu lassen, zumal für eine exakte Messung der Augenbewegung die Bewegungsfreiheit der Testpersonen eingeschränkt ist. Vor dem Hintergrund einer möglichst natürlichen Testsituation wird im Rahmen der hier durchgeführten Experimente von dieser Möglichkeit deshalb Abstand genommen, da das Hauptaugenmerk der Untersuchung eher auf Benutzerleistung und Zufriedenheit als auf dem Leseverhalten der Testteilnehmer liegt.

Allgemein ist die Methode des Laborexperiments im Kontext des IR gut erprobt, um den Einfluss einzelner Variablen auf das Nutzerverhalten zu analysieren. Dies umfasst insbesondere auch Studien die bereits in den Kapiteln 2 und 3 vorgestellt werden und die bspw. den Einfluss der Systemleistung, des Alters, des Geschlechts, der Vorerfahrung und der Aufgabenschwierigkeit untersuchen (Turpin u. Scholer, 2006; Roy u. Chi, 2003; Chevalier et al., 2015; Hölscher u. Strube, 2000; Chiou u. Wan, 2007). Als Vorteile lassen sich die Kontrollierbarkeit der Testsituation in Bezug auf Störvariablen und die damit einhergehende interne Validität der Testergebnisse hervorheben. Insbesondere ist es in der Regel nur im Labor möglich, eine einzelne Variable systematisch zu variieren, sodass Veränderungen bei einer Versuchsgruppe im Vergleich zu einer in allen wesentlichen Merkmalen gleichen Kontrollgruppe auf eben diese Variable zurückgeführt werden können. Die Replizierbarkeit der Untersuchungsbedingungen erlaubt es darüber hinaus, zur Überprüfung von Theorien oder empirisch gefundenen Zusammenhängen neue Stichproben zu generieren.

Problematisch zu sehen ist hingegen, dass die Ergebnisse von Laborstudien nicht ohne Weiteres auf den realen Suchkontext generalisierbar sind. Dazu merken Palmquist und Kim (1998, S. 15) an: „This may mean that the findings from an experiment might not fully explain the events of a real situation." So besteht die Möglichkeit, dass die Testsituation oder die Anwesenheit eines Versuchsleiters das Verhalten der Testpersonen mit beeinflusst (vgl. Abschn. 4.2.3.2). Bei der Planung eines Laborexperiments muss also darauf geachtet werden, einen möglichst hohen



(a) **(b)**

Abb. 4.1.: Schematische Darstellung von Within-Subject- und Between-Subjects-Designs. Bild (a) Within-Subject-Design: Jeder Teilnehmer in der Untersuchungsgruppe ist allen Untersuchungsbedingungen des Within-Faktors ausgesetzt (bearbeitet jede Suchaufgabe). Bild (b) Between-Subjects-Design: Jeder Teilnehmer ist nur einer Kombination der unabhängigen Variablen Systemgüte und Erwartungshaltung (Between-Faktoren) ausgesetzt.

Realitätsgrad zu erreichen, um die externe Validität der Ergebnisse sicherzustellen (vgl. Abschn. 4.1.3).

Zusammenfassend lässt sich festhalten, dass bei der Durchführung jeder experimentellen Studie zum Informationssuchverhalten das Verhältnis zwischen Kontrolle der Testsituation auf der einen und dem Realitätsgrad der Untersuchung auf der anderen Seite austariert werden muss. Vor dem Hintergrund der hier behandelten Forschungsfragen, die den konkreten Einfluss von Erwartungshaltung und Systemgüte auf Benutzerleistung und Zufriedenheit betreffen, erscheint das Laborexperiment jedoch am Besten geeignet, um das geplante Forschungsvorhaben zu verwirklichen. Dabei werden Fragen zur Balance zwischen interner und externer Validität sowie zum Anspruch eines möglichst hohen Realitätsgrads in Abschnitt 4.1.3 im Rahmen der Versuchsplanung weiter diskutiert.

4.1.2. Grundtypen experimenteller Forschungsdesigns

Im Kontext von experimentellen Studien zum Informationssuchverhalten bieten Laborstudien einen Rahmen, in dem die Wirkung weniger isolierter Einflussfaktoren auf das Nutzerverhalten untersucht werden kann. Dabei wird allgemein zwischen unabhängigen, abhängigen und Störvariablen unterschieden. Während erstere aktiv variiert werden, um ihren Einfluss auf die abhängigen Variablen zu klären, umfassen Störvariablen zusätzliche Einflussfaktoren, die nach Möglichkeit während der Untersuchung konstant gehalten bzw. durch geeignete Maßnahmen kontrolliert werden müssen. Bei mehreren unabhängigen Variablen stellt sich darüber hinaus die Frage, inwieweit sie sich in ihrer Wirkung auf die abhängigen Variablen gegenseitig beeinflussen, was als Wechselwirkung bezeichnet wird (vgl. Abschn. 4.3.2). Unterschiedliche Forschungsdesigns unterscheiden sich im Wesentlichen in der Art, wie die sich aus der Variation der unabhängigen Variablen ergebenden Untersuchungsbedingungen den Testpersonen präsentiert werden.

Grundsätzlich lassen sich, wie in Abbildung 4.1 dargestellt, zwei Typen experimenteller Forschungsdesigns für Laborstudien unterscheiden: *Within-Subject-* und *Between-Subjects-Designs*. Während im ersten Fall zwei oder mehrere Messungen mit unterschiedlichen experimentellen

Bedingungen an derselben Gruppe von Testpersonen vorgenommen werden, setzt man die Testpersonen im zweiten Fall jeweils nur einem experimentellen Stimulus aus. Üblicherweise werden in experimentellen Studien zum Informationssuchverhalten Within-Subject-Designs verwendet. So wird bspw. das System häufig als Within-Subject-Variable realisiert, wobei dieselben Versuchspersonen nacheinander zwei oder mehrere Systeme testen und die Ergebnisse hinsichtlich Sucherfolg, Zufriedenheit etc. verglichen werden. Die drei im Rahmen dieser Arbeit durchgeführten Studien werden jedoch primär als Between-Subjects-Designs durchgeführt. Der Vorteil dieses Ansatzes liegt speziell mit Blick auf die hier durchgeführten Studien darin, dass die Reaktionen der Probanden nicht durch die mehrmalige Konfrontation mit unterschiedlichen Erwartungsmanipulationen verzerrt werden, d.h. es ist weniger wahrscheinlich – wenn auch nicht ganz ausgeschlossen – dass die Probanden versuchen, die Untersuchungshypothese zu erraten und sich ihr entsprechend zu verhalten. Weitere Vor- und Nachteile beider Designs sowie beispielhafte Studien aus dem IR-Kontext sind in den folgenden beiden Abschnitten zusammengefasst.

4.1.2.1. Within-Subject-Design

Wie in der Einleitung diese Abschnitts bereits dargelegt, handelt es sich bei einem Within-Subject-Design (auch als *Messwiederholungsdesign* bezeichnet) um ein experimentelles Design, bei dem die Teilnehmer jeweils allen Variationen der unabhängigen Variablen ausgesetzt werden (vgl. Abb. 4.1). Jede einzelne Testperson stellt also eine eigene Versuchseinheit dar, deren Reaktionen auf den experimentellen Reiz (z.B. die Güte zweier IR-Systeme) erfasst und miteinander verglichen werden. Diese Methode kommt auch bei vielen Studien zum Informationssuchverhalten zum Einsatz (Allan et al., 2005; Turpin u. Scholer, 2006; Käki u. Aula, 2008; Smucker u. Jethani, 2010a; Su, 2003; Luo et al., 2017; Cox u. Fisher, 2004).

Cox und Fisher (2004) und Käki und Aula (2008) sehen insbesondere den Wegfall personenbezogener Störvariablen als entscheidenden Vorteil dieser Vorgehensweise. Käki und Aula (ebd., S. 84) argumentieren: „To control the effect of the user-related factors, we find the within-subjects design to be appropriate. In this design, all the participants use all the tested interfaces and thus, the users' personal search strategies remain constant over the tested interfaces and variation decreases." Wie an diesem Zitat deutlich wird, führt diese eingebaute Kontrolle individueller Störvariablen dazu, dass auch im Vorhinein nicht absehbare Versuchspersonencharakteristika konstant gehalten werden. Auf diese Weise werden die Ergebnisse der untersuchten Faktorstufen vergleichbarer. Gleichzeitig ist jedoch zu beachten, dass die Abhängigkeit der erhaltenen Daten bei der statistischen Auswertung berücksichtigt werden muss.

Ein weiterer Vorteil ist in der Effizienz des Designs zu sehen, da für Versuchspläne mit wiederholter Messung weniger Teilnehmer benötigt werden. Deutlich wird dies bspw. anhand der Studien von Turpin und Scholer (2006) und Allan et al. (2005). Im ersten Fall bearbeiten 30 Teilnehmer mit jedem der fünf Suchsysteme jeweils fünf Suchaufgaben, was zu einer Gesamtzahl von 1141 Suchsessions bzw. ca. 230 Sessions pro Untersuchungsbedingung führt (Turpin u. Scholer, 2006). In ähnlicher Weise generieren die 33 Versuchspersonen in der Studie von Allan et al. (2005) für acht unterschiedliche Systemgüten eine Gesamtzahl von 701 Suchsessions. Im Vergleich zu einem Between-Subjects-Design mit gleicher Teilnehmerzahl fällt die Teststärke bei einem Within-Subject-Design deshalb größer aus. Darüber hinaus lassen sich gewisse Fragestellungen, wie die zeitliche Veränderung von Merkmalen, Einstellungen und Urteilen nur

mit Hilfe dieses Ansatzes untersuchen. Zwei Beispiele für solche Fragestellungen sind in den Studien von Bruce (1994) und Tang und Solomon (2001) zu finden. Beide Studien untersuchen die Gewichtung einzelner Relevanzkriterien durch die Testteilnehmer und wie sich diese im Laufe des Suchprozesses ändert. Hier bilden also die unterschiedlichen Phasen im Suchprozess die Faktorstufen der Within-Subject-Variable. Da in diesem Zusammenhang die Veränderung der Gewichtung der Relevanzkriterien zwischen den verschiedenen Messzeitpunkten untersucht werden soll, ist es naheliegend, dass ein Vergleich zwischen verschiedenen Testpersonen nicht zielführend wäre.

Anstelle von personenbezogenen Störvariablen können in Within-Subject-Designs jedoch untersuchungsbedingte Störvariablen auftreten und die Ergebnisse beeinflussen. So ist die größere Anzahl von Suchsessions, die pro Testperson erhalten werden kann, mit einem größeren Aufwand für jede einzelne Versuchsperson verbunden. Die entsprechend verlängerten Testzeiten bergen die Gefahr von Ermüdungs- und Lerneffekten sowie allgemein einer Abnahme der Motivation der Testteilnehmer. So besteht einerseits die Möglichkeit, dass Probanden zu Beginn des Experiments eine effiziente Suchstrategie entwickeln, die bei späteren Messungen zu einer besseren Leistung beiträgt. Andererseits können Ermüdungserscheinungen im späteren Verlauf des Experiments für eine geringere Suchleistung verantwortlich sein. So wird in der bereits erwähnten Studie von Allan et al. (2005, S. 437) eine Mindesttestzeit von acht Stunden veranschlagt, die sich über zwei gesonderte Sitzungen erstreckt. Smucker und Jethani (2010a, S. 599) berichten in diesem Zusammenhang von der Problematik, dass einige Testpersonen abgeschreckt durch den Aufwand die Teilnahme an einer zweiten Testphase verweigern. Zwar können Reihenfolgeeffekte durch Randomisierung kontrolliert werden, allerdings geht dies mit einer größeren Anzahl benötigter Testpersonen einher. Auch der Umstand, dass beim Vergleich mehrerer Systeme mittels eines Within-Subject-Designs unterschiedliche Testaufgaben benötigt werden, birgt zusätzliches Störpotential, was jedoch durch weitere Randomisierung vermieden werden kann. Käki und Aula (2008, S. 84) schlagen darüber hinaus vor, nur minimale Änderungen zwischen den verwendeten Testaufgaben vorzunehmen, um eine möglichst große Vergleichbarkeit sicherzustellen: „In practice, it is often enough to just modify one term from the task description, for example: „Find Chinese restaurants in New York“ and „Find Chinese restaurants in Los Angeles“. Keeping the topics in the pairs of tasks constant is advisable as it reduces the effects that the users' interest in different topics may have on their performance.“ Ein weiterer Aspekt, der bei Within-Subject-Designs beachtet werden muss, ist die bereits in der Einleitung angesprochene Möglichkeit die forschungsleitende Untersuchungshypothese aus der Untersuchungssituation heraus zu erschließen. Durch die Konfrontation der Probanden mit allen Variationen der unabhängigen Variablen ist im Rahmen eines Within-Subject-Designs diese Möglichkeit eher gegeben. Dabei erkennen die Testteilnehmer, dass eine Änderung des Treatments stattfindet und passen in der Folge ihr Verhalten in der Weise an, wie sie denken, dass es von ihnen erwartet wird. Dieser Aspekt ist für die vorliegende Arbeit aus zwei Gründen bedeutsam: Zum einen ist vor dem Hintergrund der Forschungsergebnisse zur Wahrnehmung des eigenen Sucherfolgs (vgl. Abschn. 3.2.2) davon auszugehen, dass Versuchspersonen in der Lage sind, Qualitätsunterschiede von IR-Systemen zu erkennen und einen Systemwechsel auch ohne vorherige Ankündigung wahrnehmen können. Zum anderen erscheint es hinsichtlich der Manipulation einer kognitiven Variable, wie z.B. der

Erwartung der Testpersonen, einleuchtend, eine mehrmalige Konfrontation mit unterschiedlichen Manipulationsstufen nach Möglichkeit zu vermeiden, um kein zusätzliches Misstrauen auf Seiten der Testpersonen hervorzurufen.

Zusammenfassend lässt sich festhalten, dass Within-Subject-Versuchspläne das Potential bieten, in Bezug auf die Anzahl der Testpersonen eine möglichst große Datenbasis zu generieren. Die vor dem Hintergrund möglicher Reihenfolgeeffekte notwendigen Ausbalancierungen relativieren diesen Vorteil jedoch zu einem gewissen Grad. Nichtsdestotrotz stellt dieser Ansatz für Fragen, welche die Dynamik des Suchprozesses betreffen, das Mittel der Wahl dar, da er es ermöglicht die Reaktionen der einzelnen Versuchspersonen über den gesamten Suchverlauf hinweg zu verfolgen und zu vergleichen.

4.1.2.2. Between-Subjects Design

Im Gegensatz zum zuvor beschriebenen Within-Subject-Design werden die Teilnehmer eines Between-Subjects-Experiments nur jeweils einer Ausprägung der unabhängigen Variablen ausgesetzt (vgl. Abb. 4.1). Die Auswirkungen der unterschiedlichen Stimuli werden daher im Vergleich zwischen verschiedenen Individuen bzw. Gruppen von Individuen erfasst. Auch für diesen experimentellen Ansatz lassen sich Beispiele in der Forschungsliteratur zum Informationssuchverhalten finden (Pan et al., 2007; Tombros u. Sanderson, 1998; Smith u. Kantor, 2008; Hölscher u. Strube, 2000; Jenkins et al., 2003).

Wie bereits beschrieben werden die teilnehmenden Versuchspersonen in einzelne Untersuchungsgruppen aufgeteilt, denen jeweils eine der betrachteten Untersuchungsbedingungen zugeordnet wird. Im Vergleich zu Within-Subject-Versuchsplänen entfällt somit der Vorteil der perfekten Parallelisierung, weshalb eine Randomisierung der Untersuchungsgruppen notwendig ist, um personenbezogene Störfaktoren zu kontrollieren (vgl. Abschn. 4.2.3). Im Gegenzug verringert sich jedoch der Aufwand pro Testperson, was gerade bei Studien mit vielen Faktorstufen der unabhängigen Variablen zum Tragen kommt. In Bezug auf experimentelle Studien zum Informationssuchverhalten erlaubt dies die Bearbeitung einer größeren Anzahl von Suchaufgaben unter konstanten Untersuchungsbedingungen sowie eine anschließende Mittelwertbildung, um Topiceffekte zu reduzieren und somit die externe Validität zu erhöhen. Dieses Vorgehen findet sich bspw. in den Studien von Tombros und Sanderson (1998) und Pan et al. (2007). Des Weiteren können für jede Untersuchungsbedingung dieselben Testaufgaben verwendet werden, was sich wiederum positiv auf die interne Validität auswirkt. Dies trifft bspw. auf die Studien von Pan et al. (ebd.) und Hölscher und Strube (2000) zu.

Die Beschränkung auf einzelne Faktorstufen pro Versuchsperson führt dazu, dass die jeweiligen Messungen als unabhängig angesehen werden können, was sich vorteilhaft in Bezug auf die statistische Auswertung auswirkt. Weiterhin verringert sich die Gefahr, dass die Versuchspersonen das Untersuchungsziel durchschauen, da sie die Variationen der unabhängigen Variable nicht im Vergleich dargeboten bekommen. Dies schließt im Vergleich zu Within-Subject-Designs auch das Auftreten sog. *Halo-Effekte* aus, mit welchen die Beeinflussung der Wirkung bzw. Wahrnehmung späterer experimenteller Bedingungen durch vorangegangene Expositionen bezeichnet wird. Gerade vor dem Hintergrund, dass das dynamisierte C/D-Paradigma Halo-Effekte in Bezug auf die Erwartungshaltung antizipiert (vgl. Abschn. 3.3.1.4), erscheint dieses Vorgehen für die im Rahmen dieser Arbeit geplanten Nutzerstudien vorteilhaft.

Neben der Problematik einer höheren Varianz durch personenbezogene Störvariablen, wie Alter, Geschlecht, etc., der durch Randomisierung und Balancierung Rechnung getragen werden muss, erweist sich vor allem der zusätzliche Aufwand in Bezug auf die Stichprobengewinnung als nachteilig. Da jede Testperson im Gegensatz zum Within-Subject-Design nur Daten für eine einzige Untersuchungsbedingung generiert, werden dementsprechend mehr Testteilnehmer benötigt, um statistisch gesicherte Aussagen treffen zu können. Allerdings ist bei einigen Studien kritisch anzumerken, dass trotz allem mit einer eher geringen Anzahl von Testpersonen gearbeitet wird. So rekrutieren Hölscher und Strube (2000) lediglich sechs Versuchspersonen pro Untersuchungsgruppe, während Tombros und Sanderson (1998) sich auf zehn Teilnehmer pro Gruppe beschränken.

Zum Abschluss sei noch angemerkt, dass manche Fragestellungen sich nicht als reines Within-Subject- oder Between-Subjects-Design realisieren lassen, da sowohl unabhängige als auch Messwiederholungsvariablen berücksichtigt werden müssen. In diesem Fall bieten sich sog. *Mixed-Designs* an, die es erlauben, beide Variablentypen zu kombinieren. Ein typisches Beispiel aus dem IR-Kontext stellt die dynamische Reaktion der Versuchspersonen auf unterschiedliche Systemgüten dar. Während es sich in diesem Fall bei der Systemqualität um einen Between-Subjects-Faktor handelt, der zu unabhängigen Versuchsgruppen führt, kann die Reaktion der Testpersonen über mehrere Suchaufgaben hinweg als Messwiederholungsfaktor angesehen werden. Dieses Studiendesign wird im Rahmen dieser Arbeit bspw. für Experiment 3 gewählt (vgl. Kap. 7) und kommt auch in den Studien Smith und Kantor (2008) und Xu und Mease (2009) zum Einsatz.

Für das hier angestrebte Forschungsvorhaben lässt sich festhalten, dass sich in Bezug auf die Manipulation der Erwartungshaltung ein Between-Subjects-Design anbietet, um mögliche Austrahlungseffekte zwischen den Untersuchungsbedingungen zu vermeiden. Da insbesondere auch Wechselwirkungen zwischen Systemleistung und Erwartungshaltung betrachtet werden sollen, gilt dies analog für den Untersuchungsfaktor Systemgüte. Um darüber hinaus jedoch auch dynamische Aspekte des Suchprozesses mit in die Betrachtung einbeziehen zu können, wird dieser Ansatz in Experiment 3 um einen Messwiederholungsfaktor erweitert, der zusätzlich die Aufgabenposition berücksichtigt.

4.1.3. Planung eines Laborexperiments

Nachdem im vorherigen Abschnitt methodische Ansätze zur Analyse interaktiver Retrievalprozesse und die Einbettung spezieller Forschungsfragen in ein experimentelles Untersuchungsdesign beschrieben sind, werden im Folgenden drei grundlegende Planungsschritte diskutiert, die in gewisser Weise den methodischen Kern aller IR-Nutzerstudien bilden: die Ausgestaltung des Testsystems, die Konstruktion der Testaufgaben und das Korpusdesign. Zum einen legen sie den Kontext der simulierten Suchsituation fest und geben somit standardisierte Rahmenbedingungen für das Sucherlebnis der Testteilnehmer vor. Zum anderen bilden die hier getroffenen Entscheidungen die Grundlage für Erhebung bzw. Kontrolle der abhängigen und unabhängigen Variablen, deren Operationalisierung in Abschnitt 4.2 diskutiert wird.

4.1.3.1. Entwicklung des Testsystems

Eine zentrale Rolle im Untersuchungsdesign von Studien zum Informationssuchverhalten nimmt das gewählte Testsystem ein. Dabei sollte das Arbeiten mit dem System keine unnötige Hürde

darstellen sondern möglichst einfach, intuitiv und effizient gestaltetet sein, da das Verhalten der Testteilnehmer im Suchprozess im Vordergrund steht. Insbesondere ist darauf zu achten, dass trotz der konkreten Testsituation ein hoher Realitätsgrad erreicht wird, um ein möglichst natürliches Suchverhalten der Teilnehmer beobachten zu können.

Eine der ersten Designentscheidungen, die bei der Planung eines IR-Experiments getroffen werden muss, betrifft die Frage, ob ein real existierendes IR-System verwendet werden kann (Turpin u. Hersh, 2001) oder ob die Forschungsfrage verlangt, dass ein simuliertes Testsystem verwendet werden muss (Turpin u. Scholer, 2006; Smith u. Kantor, 2008). Letzteres ist gerade dann erforderlich, wenn eine sehr genaue Kontrolle der Systemleistung benötigt wird. Da dies für die im Rahmen dieser Arbeit durchgeführten Experimente der Fall ist, beschränkt sich die folgende Darstellung auf solche als *Wizard-of-Oz* bezeichnete Testsysteme. Um trotzdem einen einfachen Einstieg und eine gute Usability zu gewährleisten, orientieren sich die verwendeten Mockup-Systeme üblicherweise in Gestaltung und Funktionalität an den Benutzeroberflächen und Bedienkonzepten gängiger Suchmaschinen. Eine vertraute Bedienung ist nicht nur hinsichtlich der Realitätsnähe wichtig, sondern auch aus Gründen der Validität. Ein Teilnehmer, dem sich das Bedienkonzept nicht erschließt, wird unter Umständen nicht seine normale Suchleistung erreichen und die Studie somit kein valides Ergebnis erzielen.

Einen wichtigen Punkt bei der Verwendung von Mockup-Systemen stellt die Kontrolle der Systemleistung dar. Während diese in den meisten Untersuchungen durch standardisierte Evaluierungsmaße (vgl. Abschn. 4.2.1.1) bestimmt und auch manipuliert wird (Turpin u. Scholer, 2006) gibt es auch Ansätze, die andere Strategien verfolgen. Smith und Kantor (2008) bspw. simulieren unterschiedlich gute Systeme, indem je nach Untersuchungsbedingung Dokumente aus den vorderen oder hinteren Rangplätzen der entsprechenden Google-Anfrage präsentiert werden. Ein solcher Ansatz hat den Vorteil, dass die Suchergebnisse sehr aktuell sind und die Relevanzbewertung der Dokumente im Anschluss an den Test erfolgen kann. Allerdings mangelt es dieser Herangehensweise an der Möglichkeit, einen festen Systemunterschied zwischen den einzelnen Untersuchungsbedingungen vorzugeben. Ein weiterer Ansatz, bei dem zusätzlich eingefügte irrelevante Dokumente die Qualität der Ergebnisliste verringern sollen, ist problematisch, da die verschiedenen Evaluierungsmaße sehr unterschiedlich darauf reagieren (Thomas et al., 2011). Aus diesem Grund wird in dieser Arbeit der erst genannte Ansatz verfolgt, bei dem abstrakte Ergebnislisten der gewünschten Qualität im Vorhinein erzeugt und dann zur Laufzeit mit Content gefüllt werden (vgl. Abschn. 6.3.1).

Neben der Kontrolle der Systemleistung muss das Testsystem natürlich ermöglichen, den Forschungsgegenstand also das Suchverhalten der Probanden ohne Beeinträchtigungen studieren zu können. Es sollte somit den natürlichen, oftmals iterativen Suchprozess nicht behindern. Wie bereits angedeutet, muss jedoch am Ende zwischen authentischen Rahmenbedingungen auf der einen und einer Kontrolle der Einflussgrößen auf der anderen Seite abgewogen werden. Eine Reihe von Studien schränkt aus diesem Grund die Interaktion mit dem Suchsystem auf vordefinierte Suchanfragen ein (Al-Maskari et al., 2007; Huffman u. Hochster, 2007; Käki u. Aula, 2008), um allen Testpersonen die gleichen Ergebnislisten zu präsentieren. Käki und Aula (2008) favorisieren diese Methode auch, da auf diese Weise unterschiedliche Vorerfahrungen der Testteilnehmer in Bezug auf Suchthema und Rechercheerfahrung einen geringeren Einfluss auf die

Sucheffizienz ausüben. Offensichtlich geht diese Entscheidung jedoch zu Lasten eines natürlichen Sucherlebnisses der Teilnehmer.

Beispiele für Studien, die auf eigenständigen Suchanfragen basieren, sind hingegen Al-Maskari et al. (2006) und Smith und Kantor (2008). Hier werden die Anfragen direkt an eine reale Suchmaschine gerichtet und die Systemqualität a posteriori bestimmt. Wie bereits erwähnt, stellt die Kontrolle der Systemleistung einen kritischen Faktor bei diesen Studien dar. Ein dritter Ansatz beruht auf einer Synthese der beiden genannten Vorgehensweisen: Die Testpersonen können zwar eigene Suchanfragen stellen, bekommen jedoch vorgefertigte Ergebnislisten zurückgeliefert, die nicht unmittelbar auf der gestellten Suchanfrage beruhen (Turpin u. Scholer, 2006; Kelly et al., 2007). Auf diese Weise erhält man den Vorteil einer realitätsnahen Sucherfahrung zusammen mit einer guten Kontrolle der Systemleistung. Dieser Ansatz wird auch bei den im Rahmen der vorliegenden Arbeit durchgeführten Experimenten favorisiert (vgl. Abschn. 6.3.1).

Die vorangegangenen Ausführungen haben gezeigt, dass das Testsystem einen entscheidenden Einfluss auf die Güte der Benutzerstudie hat. Insbesondere bewegen sich IR-Experimente im Spannungsfeld zwischen kontrollierten Bedingungen und der Simulation eines realistischen Nutzererlebnisses. Beide Faktoren müssen gegeneinander abgewogen werden, um die richtige Balance zwischen Kontrolle der Systemleistung und freier Interaktion der Testpersonen mit dem Suchsystem zu finden. Die folgenden beiden Abschnitte beschäftigen sich mit der Erstellung der für ein *Wizard-of-Oz*-Experiment benötigten Testkollektion, welche die Wahl der Testaufgaben und den Aufbau eines entsprechenden Dokumentenkorpus umfasst.

4.1.3.2. Konstruktion der Testaufgaben

Laborexperimente mit Hilfe der im vorangegangenen Abschnitt beschriebenen Mock-Up-Systeme beruhen auf geeigneten Testkollektionen bestehend aus Suchaufgaben und zugehörigen Dokumentensammlungen. Die gewählten Testaufgaben nehmen dabei eine zentrale Rolle ein, denn sie fungieren „[...] as a proxy for the searcher's information need and establish the goals and the criteria by which search results will be assessed." (Freund u. Wildemuth, 2014, S. 1) Dieses Zitat macht deutlich, wie wichtig es ist, Suchaufgaben so zu gestalten, dass sie den natürlichen Suchprozess möglichst wenig beeinflussen. Deshalb wird bei der Konstruktion von Testaufgaben im Kontext experimenteller Studien zum Informationssuchverhalten immer nach einem Kompromiss zwischen interessanten Themen für die Zielgruppe der Probanden und der Verhinderung von Über- oder Unterforderung gesucht, um für alle Testpersonen annähernd gleiche Ausgangsvoraussetzungen zu schaffen. Ebenso wichtig sind ein ähnlicher Schwierigkeitsgrad der Aufgaben untereinander, eine gute Verständlichkeit sowie die objektive Auswertbarkeit des erreichten Sucherfolgs. Der folgende Abschnitt gibt daher einen Überblick über weitere Designentscheidungen, die im Hinblick auf die Wahl geeigneter Testaufgaben zu treffen sind. Die speziell für die einzelnen im Rahmen dieser Arbeit durchgeführten Nutzerstudien gewählten Aufgaben sind hingegen in den entsprechenden Abschnitten 5.3.4, 6.3.4 und 7.3.4 ausführlich beschrieben.

Eine erste Unterscheidung bei der Definition von Testaufgaben kann zwischen Arbeitsaufgaben (z.B. *choose and purchase a yacht*) und Suchaufgaben (z.B. *find which models of yacht are available*) getroffen werden (Toms et al., 2007, S. 360). Während Suchaufgaben sich also als zielgerichtete Aktivitäten betrachten lassen, welche die Nutzung einer Suchmaschine erfordern, ordnen Arbeitsaufgaben diese in einen übergeordneten Kontext ein (Freund u. Wildemuth, 2014,

S. 1). In diesem Zusammenhang zu nennen ist der von Borlund und Ingwersen (1997) eingeführte Ansatz der simulierten Arbeitsaufgaben (*simulated information need situation*). Hier wird der reine Suchauftrag in einen narrativen Kontext eingebettet, um eine bessere Identifikation mit der Aufgabe zu ermöglichen: „All together, the introduction of the simulated information need situation serves as a description of the 'universe' of the information need in which the test person is supposed to see him or herself, and based on which the test person formulates the search statement to the system." (ebd., S. 246) Neben den motivationalen Aspekten, die eine solche Hintergrundgeschichte ermöglicht (vgl. Abschn. 2.2), sorgt die Bereitstellung eines derartigen Narrativs für eine stärkere Uniformität im Verhalten der Testpersonen – insbesondere in Bezug auf die Relevanzbeurteilung. In diesem Sinne erlaubt es dieses Vorgehen also, das Aufgabenverständnis der Teilnehmer zu kontrollieren. Aus diesem Grund werden auch im Rahmen dieser Arbeit die einzelnen Suchaufgaben in einen narrativen Kontext eingebettet.

Eine weitere Stellschraube im Studiendesign stellen die formulierten Zielvorgaben dar. Während der überwiegende Teil interaktiver IR-Studien klassische *ad-hoc* Suchaufgaben verwendet, bei welchen die Testpersonen so viele relevante Dokumente wie möglich in einer bestimmten Zeitspanne finden sollen (Turpin u. Scholer, 2006; Smith u. Kantor, 2008) wird beim sog. *instance recall* verlangt, Dokumente mit unterschiedlichen Inhaltsaspekten ausfindig zu machen (Allan et al., 2005; Al-Maskari et al., 2006). Teilweise wird auch eine bestimmte Dokumentenanzahl vorgegeben (Kelly et al., 2007) oder eine konkret zu beantwortende Frage formuliert (z.B. *How old is the CEO of Microsoft, Bill Gates?*) (Aula, 2003). Die Wahl der Zielvorgabe erlaubt es, die Aufmerksamkeit der Testpersonen auf unterschiedliche Aspekte der realen Suchsituation zu fokussieren. Während bspw. eine ad-hoc Suchaufgabe eher zu einem explorativen Verhalten einlädt, bei dem lediglich die allgemeine Relevanz der Dokumente erfasst werden muss, erfordert ein instance recall, dass dem speziellen Informationsgehalt der Dokumente eine größere Aufmerksamkeit zuteil wird. Da der Fokus des angestrebten Forschungsvorhabens auf dem Einfluss der Systemgüte und nicht auf dem Leseverhalten der Nutzer liegt, werden im Rahmen dieser Arbeit ad-hoc Aufgaben gewählt.

Darüber hinaus lässt sich auch über die Dauer der einzelnen Aufgaben die Motivation der Testpersonen beeinflussen. Kelly (2009, S. 59) resümiert in ihrem Übersichtsartikel, dass im IIR-Bereich zwar keine allgemein anerkannten Standards für adequate Bearbeitungszeiten existieren, da diese auch vom jeweiligen Aufgabentyp und Studienziel abhängen; viele Studien den Teilnehmern aber eine Bearbeitungszeit zwischen 10 und 15 Minuten zugestehen. Als Gegenbeispiel nennt Kelly (ebd.) eine Studie von Käki und Aula (2008), in welcher das Leseverhalten der Benutzer auf Suchergebnisseiten untersucht wird. Da Käki und Aula (ebd., S. 85) in früheren Untersuchungen die Erfahrung gemacht haben, dass eine zu großzügige Zeitvorgabe bei der Relevanzbewertung dazu führen kann, dass Probanden in der Testsituation besonders gründlich vorgehen, beschränken sie die Bearbeitungszeit auf eine Minute pro Aufgabe. Wie Käki und Aula (ebd.) auch selbst bemerken, führt solch eine extreme Zeitbeschränkung zu einem größeren Zeitdruck für die Testteilnehmer. Tatsächlich verwenden Chiou und Wan (2007) die zugestandene Bearbeitungszeit, um den Schwierigkeitsgrad der Aufgaben zu variieren (vgl. Abschn. 2.4). Anderen Studien überlassen die Bearbeitungszeit dem eigenen Ermessen der Testpersonen, um möglichst natürliche Bedingungen schaffen zu können (Allan et al., 2005; Kelly et al., 2007;

Smith u. Kantor, 2008). In dieser Arbeit wird in Bezug auf die Zeitvorgabe ein gemischter Ansatz favorisiert, bei dem eine maximale Zeitvorgabe mit der Möglichkeit einer früheren Beendigung durch die Versuchspersonen kombiniert wird.

Neben diesen stärker auf die Präsentation der Aufgaben abzielenden Aspekten ist selbstverständlich auch ihre inhaltliche Ausrichtung und insbesondere der wahrgenommene Schwierigkeitsgrad von entscheidender Bedeutung (vgl. Abschn. 2.4). In einer Metastudie werten Wildemuth et al. (2014) über 100 IR-Experimente aus und gelangen zu dem Schluss, dass sich die Komplexität einer Aufgabe anhand von objektiven Kriterien (z.B. Anzahl Unteraufgaben/Facetten/Quellen) bestimmen lässt, während die Aufgabenschwierigkeit stärker von der individuellen Wahrnehmung der Suchsituation abhängt (z.B. eigene Suchleistung, Anteil relevanter Dokumente in der Suchergebnisliste, übereinstimmende Terminologie in Aufgabenbeschreibung u. Suchergebnissen) (ebd.). Weitere Studien zu diesem Thema sind in Abschnitt 2.4 beschrieben. Diese machen deutlich, dass der wahrgenommene Schwierigkeitsgrad einer Suchaufgabe nicht allein von den verwendeten Informationsbedürfnissen abhängt, sondern gleichermaßen durch den Gesamtkontext der Suche bestimmt wird. Deshalb können an dieser Stelle keine konkreten Handlungsempfehlungen abgeleitet werden. Um eine Vergleichbarkeit zwischen verschiedenen Studien zu ermöglichen, kann jedoch z.B. auf das *Repository of Assigned Search Tasks (RepAST)* zurückgegriffen werden (Freund u. Wildemuth, 2014). Auch eine Orientierung an den in Evaluierungsinitiativen wie CLEF und TREC verwendeten Suchaufgaben bietet sich an, da es sich hier um bereits erprobte Testaufgaben handelt. Darüber hinaus sollten die geplanten Aufgaben in einem Pretest auf Verständlichkeit und Akzeptanz unter den Zielpersonen überprüft werden. Bei mehreren Testaufgaben ist außerdem sicherzustellen, dass der Schwierigkeitsgrad der Aufgaben nicht zu unterschiedlich ausfällt. Dies dient zum einen dazu, dass die beobachteten Unterschiede wirklich auf die unabhängigen Variablen zurückgeführt werden können und somit das Auftreten sog. Topiceffekte zu vermeiden. Zum anderen könnten einzelne als zu schwierig empfundene Aufgaben Auswirkungen auf die Selbstwirksamkeitserwartung ausüben und damit den Einfluss der Erwartungshaltung in Bezug auf nachfolgende Aufgaben verfälschen (vgl. Abschn. 2.2.2). Für das hier angestrebte Forschungsvorhaben wird deshalb darauf geachtet, gemeinhin bekannte Themen zu wählen, um so für alle Testpersonen annähernd gleiche Ausgangsvoraussetzungen zu gewährleisten.

Die vorangegangene Diskussion macht die grundlegende Bedeutung der Auswahl der Suchaufgaben für das Design eines IR-Experiments deutlich und stellt wesentliche für die Nutzerstudien getroffene Designentscheidungen dar. Der Fokus liegt dabei auf Aufgaben, die sich möglichst nah an der alltäglichen Suchsituation durchschnittlicher Internetnutzer orientieren. Nach diesem Überblick über den Einfluss der Aufgabenauswahl auf die Testergebnisse wird im nächsten Abschnitt genauer auf die Entwicklung eines Goldstandards für den Vergleich des Suchverhaltens verschiedener Benutzer eingegangen.

4.1.3.3. Korpusdesign und Annotation

Ein weiterer Aspekt, der bei der Wahl der Testaufgaben zu berücksichtigen ist, betrifft die Bereitstellung einer ausreichenden Dokumentenbasis, die im Rahmen der Nutzerstudie durchsucht werden kann. Wird dabei nicht auf ein reales Suchsystem zurückgegriffen, muss solch eine Testkollektion im Vorhinein erstellt werden. Dabei werden neben der reinen Dokumentensammlung

auch objektive Relevanzbewertungen in Bezug auf die einzelnen Suchaufgaben benötigt. Wie in Abschnitt 4.1.3.1 dargelegt, dienen diese Bewertungen zum einen der Kontrolle der Systemgüte, zum anderen bilden sie die Grundlage der meisten Operationalisierungen der Benutzerleistung (vgl. Abschn. 4.2.2.2). Dieser Abschnitt fasst die im Zuge der Korpuserstellung zu treffenden Designentscheidungen sowie die übliche Vorgehensweise bei der Annotation von Dokumenten zusammen. Dabei wird zunächst auf die bereits angesprochenen Vorteile der Wiederverwendung bereits bestehender Testaufgaben sowie entsprechender Korpora eingegangen. Anschließend werden Aspekte der Skalierung der Relevanz der Dokumente sowie ihrer objektivierten Messung thematisiert, bevor schließlich die Überprüfung der Urteilerübereinstimmung, der sog. *Interrater-Reliabilität*, beschrieben wird. Das tatsächliche Vorgehen im Rahmen der einzelnen Experimente, sowie die entsprechende Verteilung relevanter und irrelevanter Dokumente innerhalb der verwendenden Testkollektionen können hingegen den Abschnitten 5.3.4, 6.3.4 und 7.3.4 entnommen werden.

Wie im vorherigen Abschnitt angedeutet, kann es aus vielerlei Hinsicht von Vorteil sein, in Bezug auf die Wahl der Testaufgaben zunächst zu prüfen, ob auf bereits bestehende Testkorpora zurückgegriffen werden kann. Neben der Zeitersparnis aufgrund des entfallenden Korpusdesigns, sind die darin enthaltenen Testaufgaben bereits erprobt und auf die Gütekriterien Objektivität, Reliabilität und Validität überprüft. Außerdem wird auf diese Weise der Vergleich mit früheren Studien erleichtert. Im Rahmen des ersten Experiments zur Übertragbarkeit des C/D-Paradigmas auf den IR-Kontext (vgl. Abschn. 5.3.4) wird deshalb dieser Ansatz verfolgt und ein bestehendes Korpus aus Pressemitteilungen verwendet. In den folgenden beiden Untersuchungen hingegen werden eigene Testkorpora auf der Basis von Internetdokumenten entwickelt. Ausgangspunkt für diese Designänderung ist die Beobachtung, dass die Aktualität der Dokumente in einem IR-Experiment stets mit bewertet wird, was das im ersten Benutzertest verwendete CLEF-Korpus aus den Jahren 2001 und 2003 für die weiteren Benutzerstudien weniger geeignet erscheinen lässt. Hinzu kommt in den letzten beiden Experimenten der Anspruch, eine möglichst realitätsnahe Umsetzung des Sucherlebnisses zu erreichen, die sich eng an der alltäglichen Erfahrung der Probanden mit Internetsuchmaschinen orientiert. In Bezug auf beide Aspekte bietet sich die Verwendung eines Webkorpus an. Neben einer hohen Aktualität der Dokumente und ihrer freien Verfügbarkeit über das Internet ist weiterhin davon auszugehen, dass alle Probanden mit Struktur und Layout von Webdokumenten vertraut sind. Nachteilig ist lediglich die Unbeständigkeit der Daten, was jedoch durch eine Speicherung der Webseiten im Vorfeld der Untersuchung umgangen werden kann.

Eine Testkollektion umfasst neben der reinen Dokumentensammlung auch objektive Relevanzbewertungen in Bezug auf die einzelnen Suchaufgaben. Diese Bewertungen erfolgen durch Juroren, die, um die Objektivität der Relevanzurteile zu gewährleisten, in der Regel randomisierte Listen von Dokumenten erhalten und diese entsprechend ihrer Relevanz annotieren. Die Annotation kann je nach Untersuchungsziel anhand unterschiedlich gestufter Ratingskalen erfolgen. Während nach wie vor viele Autoren davon ausgehen, dass Relevanz typischerweise binär ist und dementsprechend dichotome Skalen verwenden (Tombros u. Sanderson, 1998; Belkin et al., 1999; Hersh et al., 2000; Allan et al., 2005; Al-Maskari et al., 2006; Kelly et al., 2007; Smucker u. Jethani, 2010a), unternehmen andere Forscher den Versuch einer genaueren Messung und

Tab. 4.1.: Kreuztabellen für die Berechnung von Cohens Kappa. Dargestellt sind die Kreuztabellen der relativen Häufigkeiten (H) der einzelnen Bewertungskategoriepaarungen für die binäre und 4-stufige Relevanzskala. H_{eI_1/R_2} bezeichnet bspw. den Anteil der Dokumente, den Juror 1 als eher irrelevant und Juror 2 als relevant bewertet. Der in Gleichung (4.1) mit p bezeichnete Anteil der übereinstimmenden Bewertungen ergibt sich als Summe der Diagonaleinträge der Tabelle. p_e hingegen erhält man als Summe der Produkte der Spalten- und Zeilensummen (SS u. ZS) ($p_e = \sum_i Z_i \cdot S_i$).

		Juror 1		
		rel.	**irr.**	**ZS**
Juror 2	**rel.**	H_{R_1/R_2}	H_{I_1/R_2}	Z_1
	irr.	H_{R_1/I_2}	H_{I_1/I_2}	Z_2
	SS	S_1	S_2	

		Juror 1				
		rel.	**eher rel.**	**eher irr.**	**irr.**	**ZS**
Juror 2	**rel.**	H_{R_1/R_2}	H_{eR_1/R_2}	H_{eI_1/R_2}	H_{I_1/R_2}	Z_1
	eher rel.	H_{R_1/eR_2}	H_{eR_1/eR_2}	H_{eI_1/eR_2}	H_{I_1/eR_2}	Z_2
	eher irr.	H_{R_1/eI_2}	H_{eR_1/eI_2}	H_{eI_1/eI_2}	H_{I_1/eI_2}	Z_3
	irr.	H_{R_1/I_2}	H_{eR_1/I_2}	H_{eI_1/I_2}	H_{I_1/I_2}	Z_4
	SS	S_1	S_2	S_3	S_4	

Analyse von Relevanz, indem sie den Juroren eine differenziertere Skala mit mehr Ausprägungen zur Verfügung stellen (Spink u. Greisdorf, 2001; Sormunen, 2002; Purgailis Parker u. Johnson, 1990; Smith u. Kantor, 2008). Eine ausführlichere Diskussion verschiedener im IR-Kontext bereits verwendeter Verfahren findet sich in Abschnitt 4.2.2.1. Während im zweiten Experiment ebenfalls eine dichotome Relevanzskala zugrunde gelegt wird, bewerten die Juroren die Dokumente für das dritte Experiment auf einer 4-stufigen Relevanzskala. Dies erlaubt sowohl eine feinere Abstimmung der Systemgüte der präsentierten Ergebnislisten als auch eine detailliertere Analyse der Benutzerleistung (vgl. Abschn. 4.2.2.2). So kann bspw. direkt überprüft werden, ob Nutzer eines besseren Systems nur noch besonders relevante Dokumente als relevant akzeptieren.

Während im Rahmen von TREC die Relevanz der Dokumente als „relevant to a single user at a single point in time“ begriffen wird und somit nicht ohne Weiteres auf die Relevanzwahrnehmung anderer Nutzer übertragbar ist (Kelly, 2009, S. 72), wird im Rahmen dieser Arbeit ein anderer Ansatz verfolgt. Um konsistente und über den einzelnen Juror hinaus generalisierbare Relevanzbewertungen zu erhalten, werden pro Dokument mehrere Jurorenurteile eingeholt und anschließend das Relevanzniveau per Mehrheitsentscheid festgelegt. Dabei wird den Juroren für die binäre Relevanzskala des zweiten Experiments keine bestimmte Bewertungsmethode vorgegeben, sondern lediglich die Suchaufgabe präsentiert, um ein möglichst natürliches Bewertungsszenario nachzubilden. Für die im Rahmen des dritten Experiments verwendete 4-stufige Bewertungsskala werden hingegen explizite Bewertungskategorien für die Abstufungen *relevant*, *eher relevant*, *eher irrelevant* und *irrelevant* vorgegeben, die sich an der Skala von Sormunen (2000, S. 63) orientieren. Auf diese Weise soll auch für diese komplexere Relevanzskala eine möglichst konsistente Klassifikation der Dokumente im Testkorpus ermöglicht werden. Demgegenüber steht im Rahmen der Relevanzbewertung durch die Probanden im Verlauf des Experiments die Unterstützung einer möglichst intuitiven Relevanzklassifikation im Vordergrund, sodass in diesem Fall auf eine von Tang et al. (1999) und Moosbrugger und Kelava (2011) als optimal empfohlene Skala mit acht Abstufungen zurückgegriffen wird (vgl. Abschn. 4.2.2.1). Diese Aufteilung in eine Juroren- und Teilnehmerskala hat den Vorteil, möglichst konsistente, auf klar definierten Relevanzkategorien beruhende, Bewertungen für das Testkorpus zu erhalten, während gleichzeitig den Teilnehmern feinere Abstufungen für ihre Bewertung zur Verfügung stehen, die jedoch leicht durch das Zusammenfassen von je zwei Bewertungskategorien in eine

äquivalente 4-stufige Skala transformierbar sind.

Um die Konsistenz der Jurorenurteile zu überprüfen, könnte prinzipiell der Anteil der Dokumente betrachtet werden, den zwei Juroren übereinstimmend als relevant bzw. irrelevant bewerten. Dieses einfache Gütekriterium hat jedoch den Nachteil, dass es die Wahrscheinlichkeit zufälliger Übereinstimmungen nicht berücksichtigt, die umso höher ausfällt, je weniger Bewertungskategorien vorhanden sind (50% bei binären Skalen) (Bortz u. Döring, 2006, S. 272). Aus diesem Grund wird in der Literatur stattdessen die Interrater-Reliabilität anhand von Cohens Kappa bevorzugt (Cohen, 1960). Diese Kennzahl korrigiert den Anteil der von zwei Juroren übereinstimmend als relevant bzw. irrelevant bewerteten Dokumente gerade um die Wahrscheinlichkeit zufälliger Übereinstimmungen (Bortz u. Döring, 2006, S. 272 f.). Zur Berechnung von Cohens Kappa wird zunächst eine Kreuztabelle mit den relativen Häufigkeiten aller Bewertungskategoriepaarungen erstellt (vgl. Tab. 4.1). Ausgehend von diesen Werten ergibt sich Cohens Kappa dann als (ebd., S. 272 f.)

$$\kappa_{\text{Cohen}} = \frac{p - p_e}{1 - p_e}\,, \tag{4.1}$$

wobei p dem Anteil der Übereinstimmungen, d.h. der Summe der Diagonaleinträge der Kreuztabelle entspricht. Die Größe p_e stellt einen Schätzwert für die zufälligen Übereinstimmungen dar und berechnet sich als die Summe der Produkte der jeweiligen Spalten- und Zeilensummen ($p_e = \sum_i Z_i \cdot S_i$). Je höher der Wert von Cohens Kappa ausfällt, desto höher ist die Übereinstimmung zwischen den Juroren. Zur Einschätzung der Übereinstimmungsstärke sind in der Literatur folgende Leitlinien zu finden. Werte zwischen $0{,}6$ und $0{,}75$ zeigen eine gute Übereinstimmung (ebd., S. 273) während Werte zwischen $0{,}4$ und $0{,}6$ als ausreichend gelten (Greve u. Wentura, 1997, S. 111).

Im Ergebnis stellen die vorangegangenen Ausführungen einen Leitfaden für die Erstellung neuer Testkollektionen dar. Seine Anwendung auf die konkret gewählten Testaufgaben für die Experimente 2 und 3 ist in den Abschnitten 6.3.4 und 7.3.4 näher ausgeführt. In Bezug auf die Planungsschritte für eine experimentelle Studie zum Informationssuchverhalten verbleiben somit nur noch die Operationalisierung der abhängigen und unabhängigen Variablen sowie die Beschreibung der Auswertungsmethodik, die im verbleibenden Teil dieses Kapitels erläutert werden.

4.2. Operationalisierung und Messung der untersuchten Variablen

Nachdem das Hauptaugenmerk beim Stand der Forschung auf dem Einfluss der betrachteten Variablen auf die Suchergebniswahrnehmung liegt, werden in diesem Abschnitt die verschiedenen methodischen Herangehensweisen zur Operationalisierung der in dieser Arbeit untersuchten Variablen zusammengetragen und diskutiert. Auf die tatsächliche Umsetzung wird zur besseren Nachvollziehbarkeit erst im Zuge der Durchführung der einzelnen Experimente in den Kapiteln 5 bis 7 eingegangen. Eine Definition des Variablenbegriffs findet sich bspw. bei Kelly (2009, S. 35): „Variables are present in almost all studies. Variables represent concepts. Specifically they represent ways of defining, observing and measuring the concepts that researchers aim to study. Relevance, performance, and satisfaction are all concepts.“ Im Rahmen eines Untersuchungsdesigns unterscheidet man zwischen abhängigen und unabhängigen Variablen. Dabei stellen

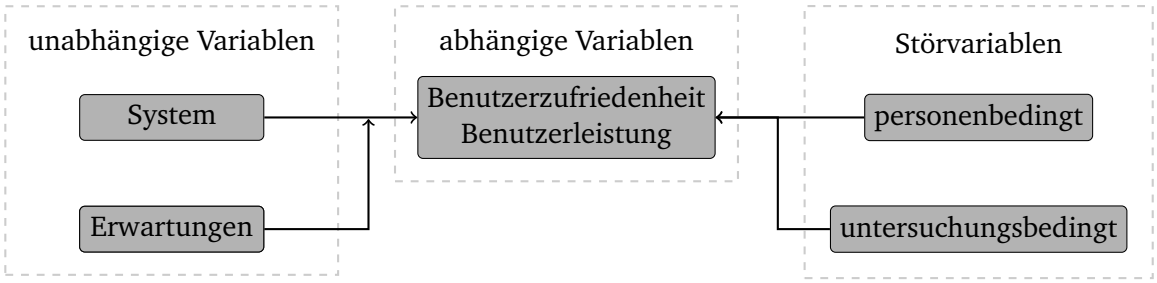


Abb. 4.2.: Theoretisches Untersuchungsmodell und Klassifizierung globaler Variablen.

die abhängigen Variablen die interessierenden Messgrößen dar, deren Abhängigkeiten von den unabhängigen Variablen, die unter der Kontrolle des Forschers stehen, aufgeklärt werden sollen (vgl. Kap. 3). Von den unabhängigen Variablen, die im Experiment zur Erklärung der abhängigen Variablen herangezogen werden, unterscheidet man sog. *Störfaktoren*, die einen zusätzlichen Einfluss auf die Untersuchungsergebnisse haben können (vgl. Kap. 2) und somit experimentell kontrolliert werden müssen.

Zur Erinnerung zeigt Abbildung 4.2 noch einmal das im Rahmen dieser Arbeit verwendete Untersuchungsmodell mit den entsprechenden Dependenzstrukturen. Dabei stellen Systemgüte und Erwartungshaltung die unabhängigen Variablen dar, die aktiv manipuliert werden. Gemessen hingegen wird der wechselseitige Einfluss dieser beiden Faktoren auf Benutzerzufriedenheit und -leistung, die somit die unabhängigen Variablen darstellen, während weitere Personenvariablen statistisch kontrolliert werden, um Konfundierungen zu vermeiden. Gemäß dieses Untersuchungsmodells wird im Folgenden zunächst die Manipulation der unabhängigen Variablen erläutert. Auf die Messung der abhängigen Variablen sowie die verwendeten Strategien zur Kontrolle zusätzlicher Einflussgrößen wird in den beiden darauf folgenden Unterabschnitten eingegangen.

4.2.1. Unabhängige Variablen der Benutzer-System-Interaktion

In allen drei dieser Arbeit zugrundeliegenden Nutzerstudien stellen die Systemgüte und die Erwartungshaltung der Testteilnehmer die unabhängigen Variablen dar, die im Rahmen der Experimente variiert werden. Der folgende Abschnitt diskutiert für beide Faktoren Ansätze zur Manipulation dieser Größen.

4.2.1.1. Manipulation der Systemleistung

Die Manipulation der Systemleistung zur Klärung des Einflusses der Systemgüte auf das Benutzerverhalten hat eine lange Tradition in der IIR-Forschung. Bevor jedoch die Methoden zu ihrer Manipulation diskutiert werden können, ist es zunächst notwendig, den Begriff der Systemleistung und ihre Quantifizierung darzustellen. Die Frage nach dem passenden Maß, um die Güte unterschiedlicher Suchsysteme vergleichbar zu machen, stellt sich bereits im Kontext der systemzentrierten IR-Evaluierung und hat früh zu den beiden Standardmaßen *Recall* und *Precision* geführt. Wie in Abschnitt 3.2.1 im Rahmen der Übertragbarkeit systemorientierter Ergebnisse auf den Nutzerkontext jedoch bereits deutlich wird, haben die beobachteten Kompensationseffekte bzw. fehlenden Korrelationen mit diesen klassischen Evaluierungsmaßen zur Entwicklung einer Reihe von alternativen Bewertungsstrategien geführt, die versuchen, die Nutzerperspek-

Tab. 4.2.: Übersicht über in der Literatur verwendete Systemleistungsmaße. Um die Übersichtlichkeit der Tabelle zu gewährleisten, werden in den Formeln die folgenden Abkürzungen verwendet: Anzahl der Dokumente in einer Ergebnisliste (L), Anzahl relevanter bzw. irrelevanter Dokumente auf den ersten n Rankingpositionen ($R(n)$ bzw. $I(n)$). $rel(n)$ ist 1 wenn das Dokument an Position n in der Ergebnisliste relevant ist und 0, wenn es irrelevant oder nicht bewertet ist. Die Darstellung orientiert sich an Kelly (2009, S. 109 ff.), erweitert um die entsprechende Darstellung als mathematische Formel.

Maß	**Beschreibung**	**Formel**
Precision (Prec)	Anteil relevanter Dokumente in Ergebnisliste.	$\frac{R(\mathrm{L})}{L}$
Recall (Rec)	Anteil gefundener relevanter Dokumente.	$\frac{R(L)}{R(\mathrm{Korpus})}$
Precision at n ($P@n$)	Precision der Ergebnisliste bis zur einschließlich n-ten Listenposition.	$\frac{R(n)}{n}$
Average Precision (AvP)	Mittelwert der $P@n$ über alle relevanten Dokumente in der Ergebnisliste.	$\sum_{n=1}^{L} \frac{rel(n) \cdot P@n}{R(L)}$
Mean Average Precision (MAP)	Mittelwert der AvP über T verschiedene Suchanfragen.	$\sum_{t=1}^{T} \frac{AvP(\mathrm{Liste}_t)}{T}$
Binary Preference (BPref)	Zählt wie oft im Mittel relevante vor irrelevanten Dokumenten zurückgegeben werden. Hierbei werden nicht bewertete Dokumente in der Ergebnisliste ignoriert. In diesem Fall kann also $I(\mathrm{Liste}) + R(\mathrm{Liste}) < L$ gelten.	$\sum_{n=1}^{L} \frac{rel(n)}{R(L)} \cdot \left(1 - \frac{I(n)}{\min(R(L), I(L))}\right)$
Cumulative Gain (CG)	Summe der Relevanzwerte aller zurückgegebenen Dokumente. Bei einer erweiterten Relevanzskala kann $rel(n)$ auch andere Werte als 0 und 1 annehmen.	$\sum_{n=1}^{L} rel(n)$
Discounted Cumulative Gain (DCG)	Summe der Relevanzwerte aller zurückgegebenen Dokumente. Im Gegensatz zu CG wird jedoch die jeweilige Relevanz eines Dokuments mit dem Logarithmus seiner Listenposition gewichtet, um den Aufwand für den Nutzer zum Auffinden des Dokuments zu berücksichtigen.	$\sum_{n=1}^{L} \frac{rel(n)}{\log_2(n+1)}$
normalized Discounted Cumulative Gain (nDCG)	Hier wird der gegebene DCG-Wert einer Ergebnisliste mit dem maximalen DCG-Wert normalisiert ($nDCG_{ideal}$), der bei optimaler Sortierung der Dokumente in der Ergebnisliste erreicht werden könnte.	$\frac{nDCG}{nDCG_{ideal}}$

tive stärker zu berücksichtigen. Eine Übersicht über eine Reihe von Systemleistungsmaßen ist in Tabelle 4.2 zusammengefasst. Im Gegensatz zu den klassischen Maßen Recall und Precision, die alle Rankingpositionen gleichberechtigt in die Bewertung der Retrievalleistung einbeziehen, berücksichtigt die Mehrheit der neueren Leistungsmaße auch die relative Position relevanter Dokumente in der Ergebnisliste. Dahinter steht zum einen die Idee, auch dem Aufwand Rechnung zu tragen, den ein Nutzer zum Auffinden der relevanten Dokumente betreiben muss. Zum anderen zeigen Nutzerstudien, dass Suchmaschinennutzer tatsächlich dazu neigen, überhaupt nur die ersten zehn Treffer in der Ergebnisliste zu betrachten (Jansen et al., 2000), bevor sie eine neue Suchanfrage stellen oder ihre Suche beenden. Die verschiedenen Leistungsmaße unterscheiden sich dann im Wesentlichen durch die gewählte Gewichtung im Hinblick auf die Rankingpositionen.

Im Fall der *Average Precision* (AvP) wird bspw. für jedes relevante Dokument in der Ergebnisliste die Precision der Liste bis zu dieser Rankingposition ($P@n$) berechnet und über diese anschließend der Mittelwert gebildet. Dies kann auch als Flächeninhalt unter der Recall-Precision-Kurve interpretiert werden. In vielen Fällen wird die auf der AvP beruhende *Mean Average Precision*

(MAP) betrachtet. Zu ihrer Berechnung wird über die AvP verschiedener Suchanfragen der Mittelwert gebildet. Diese Berücksichtigung mehrerer Suchanfragen ermöglicht es, unterschiedliche Benutzerstandpunkte zu integrieren. Beiden Maßen ist gemein, dass sie als etabliert und zuverlässig gelten. Problematisch ist jedoch die Tatsache, dass sehr unterschiedliche Verteilungen der relevanten Dokumente auf den obersten Listenplätzen zu denselben AvP-Werten führen können. Einem typischen Benutzer, der überhaupt nur die ersten 10 Listenplätze einer langen Ergebnisliste betrachtet, können zwei Listen mit identischem AvP-Wert also sehr unterschiedlich erscheinen.

Die Notwendigkeit zur Erstellung umfangreicher Testkorpora zur Evaluierung von Suchsystemen geht mit einem erheblichen Aufwand für die Relevanzbewertung der enthaltenen Dokumente einher. Eine Strategie, um diesen Aufwand bei systembasierten IR-Evaluierungen zu reduzieren, ist die sog. Pooling-Methode, bei der nur ein Teil der von den Systemen zurückgelieferten Dokumente tatsächlich von Juroren bewertet wird (Womser-Hacker, 2004, S. 229). Aus diesem Grund wächst das Interesse an Systemleistungsmaßen, die robust auf das Vorhandensein von nicht evaluierten Dokumenten reagieren (Moghadasi et al., 2013). Im Kontext der TREC-Evaluierungsinitiative hat sich dabei in den letzten Jahren die von Buckley und Voorhees (2004) eingeführte *Binary Preference* (BPref) etabliert. Diese kann auch auf Ergebnislisten angewendet werden, die Dokumente enthalten für die keine Jurorenurteile vorliegen, was gerade bei großen Dokumentensammlungen von Vorteil ist, weil sich dadurch der Bewertungsaufwand reduzieren lässt. Für die Berechnung werden dann einzig Informationsobjekte mit bekannter Relevanz berücksichtigt. Dabei misst die BPref im Wesentlichen, wie häufig relevante vor irrelevanten Dokumenten gerankt werden (vgl. Tab. 4.2). Während also im Fall der AvP nicht bewertete Treffer als irrelevant betrachtet werden, berücksichtigt dieses Maß ausschließlich bewertete Dokumente. Buckley und Voorhees (ebd.) berichten, dass sich BPref robust in Bezug auf nicht bewertete Dokumente innerhalb der Ergebnisliste verhält. Neue Studien von Sakai (2007) und Sakai und Kando (2008) zeigen jedoch, dass sich ähnliche Ergebnisse auch für AvP oder nDCG erreichen lassen, wenn diese für reduzierte Ergebnislisten, die lediglich die Dokumente mit bekannter Relevanzbewertung enthalten, ausgewertet werden.

Die sog. *Cumulative Gain Maße* hingegen gehen davon aus, dass ein Benutzer einen zusätzlichen Nutzen aus jedem weiteren relevanten Dokument zieht (Järvelin u. Kekäläinen, 2000; Järvelin u. Kekäläinen, 2002). Diesen Nutzen versuchen Maße wie *Cumulative Gain* (CG) zu operationalisieren, indem sie den jeweiligen Gewinn in Form der Relevanz der Informationsobjekte aufsummieren. Dies erlaubt insbesondere auch die Verwendung nicht-binärer Relevanzskalen. Der *Discounted Cumulative Gain* (DCG) berücksichtigt darüber hinaus auch die Rankingposition als Maß für den Aufwand der zum Auffinden eines Informationsobjekts erbracht werden muss, indem die Relevanz durch den Logarithmus der Rankingposition geteilt wird. Somit wird einem relevanten Dokument an Position zehn ein niedrigerer Relevanzwert zugeordnet als wenn es an Position fünf gestanden hätte. Der *normalized Discounted Cumulative Gain* (nDCG) schließlich, normalisiert den Gewinn an einem hypothetischen, optimalen Ranking, welches alle relevanten Dokumente nach oben stellt. Auf diese Weise werden auch Listen verschiedener Länge vergleichbar (Järvelin u. Kekäläinen, 2002).

Nach dieser Übersicht über die gebräuchlichsten in der Literatur verwendeten Systemleistungsmaße wird im Folgenden näher auf die Manipulation dieser Variable im Kontext von experi-

mentellen Studien zum Informationssuchverhalten eingegangen. Der Vergleich verschiedener Suchsysteme anhand unterschiedlicher Leistungsmaße stellt den Kern der Cranfield-Methode zur Evaluation von Suchmaschinen dar. Die entsprechende Übertragbarkeit dieser Ergebnisse wird in Abschnitt 3.2.1 behandelt. Dort werden eine Reihe von Studien besprochen, welche die Systemleistung als unabhängige Variable variieren, indem Ergebnislisten mit unterschiedlichen Leistungswerten in Bezug auf eines der Evaluierungsmaße präsentiert werden (Turpin u. Scholer, 2006; Smith u. Kantor, 2008; Smucker u. Jethani, 2010a; Allan et al., 2005). Diese Strategie wird auch im Rahmen dieser Arbeit verfolgt. Die Wahl fällt dabei auf die AvP als im Kontext von Nutzerstudien etabliertes Leistungsmaß, das auch in einer Reihe anderer Studien zum Einsatz kommt. Vor dem Hintergrund der gemischten Ergebnisse in Bezug auf die Übertragbarkeit auf den Nutzerkontext, wird jedoch darauf geachtet, dass die verwendeten Ergebnislisten auch einen Systemunterschied in Bezug auf die anderen in diesem Abschnitt vorgestellten Leistungsmaße aufweisen (vgl. Abschn. 6.3.1). Um die Anzahl der Untersuchungsgruppen in Kombination mit der Manipulation der Erwartungshaltung nicht zu groß werden zu lassen, wird darüber hinaus entschieden, die Studie auf zwei unterschiedliche Systemgüten zu beschränken. Für die tatsächlichen Systemunterschiede wird in Bezug auf die AvP ein mittleres Leistungsniveau von $0{,}55$ für das schlechtere und ein im oberen Drittel angesiedeltes Leistungsniveau von $0{,}75$ für das bessere System gewählt, was einem relativen Systemunterschied von $0{,}35\%$ entspricht. Die gewählten Systemunterschiede fallen dabei in das Untersuchungsspektrum vergleichbarer experimenteller Studien wie bspw. Allan et al. (2005) und Turpin und Scholer (2006). Durch die Wahl eines Mock-Up-Designs für das Testsystem (vgl. Abschn. 4.1.3.1) lässt sich darüber hinaus der gewählte Systemunterschied vollständig kontrollieren, da Ergebnislisten mit den exakten AvP-Werten im Vorfeld der Untersuchung erzeugt werden können. Im Zuge der erweiterten Relevanzskala für Experiment 3 werden bei der randomisierten Zuteilung der Dokumente zu den Rankingplätzen jeweils zur Hälfte relevante und eher relevante Dokumente bzw. zur Hälfte irrelevante und eher irrelevante Dokumente verwendet.

Nach dieser Übersicht über die in der Literatur verwendeten Systemleistungsmaße und der Diskussion der für die Untersuchungen gewählten Systemleistungsunterschiede werden im folgenden Abschnitt Verfahren zur Manipulation der Erwartungshaltung vorgestellt.

4.2.1.2. Manipulation der Erwartungshaltung

Im Gegensatz zur Systemleistung spielt die Erwartungshaltung der Nutzer in der IR-Forschung bisher nur eine geringe Rolle. Zwar postuliert bspw. Kelly (2009, S. 38 f.) einen moderierenden Einfluss der Erwartungshaltung auf die Nutzerzufriedenheit, insbesondere aber mögliche Wechselwirkungen mit der Systemleistung sind bisher kaum erforscht und stehen deshalb im Zentrum des hier angestrebten Forschungsvorhabens. Analog zur Systemleistung soll auch die Erwartungshaltung im Rahmen der Untersuchungen aktiv manipuliert werden. Im Folgenden werden zunächst kurz die Ansätze vierer Studien diskutiert, die im weitesten Sinne Erwartungen berücksichtigen. Da diese Untersuchungen bereits im Kontext der Benutzerzufriedenheit in Abschnitt 3.3.2.1 ausführlich vorgestellt sind, beschränkt sich die folgende Darstellung auf das jeweilige Vorgehen zur Erwartungsmanipulation. Im Anschluss daran werden die im Rahmen dieser Arbeit gewählten Manipulationsstrategien erläutert.

Szajna und Scamell (1993) beschäftigen sich mit dem Einfluss von Erwartungen bei der Nut-

zung betrieblicher Informationssysteme. Die Manipulation der Erwartungshaltung erfolgt in drei Stufen (niedrig, moderat u. hoch) mit Hilfe von Instruktionstexten im Rahmen einer ersten Testsession, an die sich drei weitere Sessions für die tatsächliche Aufgabenbearbeitung anschließen. Der Manipulationserfolg wird zu Beginn der ersten Aufgabensession durch einen Fragebogen überprüft und eine ANOVA ergibt hier einen signifikanten Unterschied zwischen allen drei Treatmentgruppen. Des Weiteren können Szajna und Scamell (1993) einen signifikanten Einfluss der Erwartungsmanipulation auf die Benutzerzufriedenheit nachweisen. In einem erweiterten Sinne untersuchen auch Cuadra und Katter (1967, S. 297 ff.) den Einfluss von Instruktionstexten auf die Erwartungshaltung. In einer Laborstudie erfassen sie in diesem Zusammenhang den Einfluss von vierzehn unterschiedlichen Nutzungskontexten auf die Relevanzbeurteilung von 140 Juroren (vgl. Abschn. 2.1.1.3). Beispiele für die in der Studie berücksichtigten Kontexte sind das Erstellen einer Literaturliste für Studenten, das Verfassen eines Reviewartikels oder ein spezielles Interesse an den in verschiedenen Studien verwendeten Methoden. Die Ergebnisse deuten darauf hin, dass der von den Juroren wahrgenommene Relevanzgehalt von Suchergebnislisten tatsächlich von dem präsentierten Nutzungsziel abhängig ist (ebd., S. 302). Einen ähnlichen Ansatz verfolgen auch Jansen et al. (2007), um den Einfluss von Markenwahrnehmungen auf die Relevanzbeurteilung von Suchergebnissen zu analysieren. Die Erwartungsmanipulation erfolgt hier durch die Präsentation identischer Trefferlisten als Suchergebnis unterschiedlicher Suchmaschinen (Google, MSN, Yahoo u. ein unbekanntes System). Auch hier zeigt sich ein deutlicher Einfluss auf die Bewertung der Suchmaschinenqualität, wobei im Vergleich die unbekannte Suchmaschine von den Teilnehmern am schlechtesten bewertet wird. Der Einfluss übertroffener bzw. enttäuschter Erwartungen hingegen wird in einer Studie von Cox und Fisher (2004) untersucht. Hier erfolgt die Manipulation der Erwartungshaltung über die Formulierung der Testaufgaben bzw. die Aufgabenschwierigkeit. Die Manipulation wird mit Hilfe eine Fragebogens überprüft. Im Rahmen der Studie kann gezeigt werden, dass die Differenz zwischen gemessener Ausgangserwartung und Leistungsbewertung der präsentierten Ergebnisliste positiv mit der Nutzerzufriedenheit korreliert. Im Gegensatz zu dem hier angestrebten Forschungsvorhaben, bei dem die Systemleistung unabhängig vom Benutzer definiert ist, wird hier also ausschließlich die Benutzersicht berücksichtigt.

Nach dieser Übersicht über methodische Ansätze der Erwartungsmanipulation, wird nun die im Rahmen dieser Arbeit verfolgte Strategie vorgestellt. Dabei wird von einem Vorgehen wie bei Cox und Fisher (ebd.) abgesehen, da hier sehr speziell die mit der Aufgabenschwierigkeit verknüpfte Erwartung zum Tragen kommt, die nicht unbedingt von der Systemleistung abhängt. Vielmehr sind die Aufgaben bei Cox und Fisher (ebd.) so formuliert, dass es den Teilnehmern bei niedriger Erwartungshaltung allgemein fraglich erscheint, die entsprechende Aufgabe mit Hilfe einer Suchmaschine lösen zu können. Darüber hinaus berichten Cox und Fisher (ebd.) von einer weiteren Tücke dieses Ansatzes, der mit dem Vorwissen der Testpersonen zu tun hat (vgl. Abschn. 2.1.1.3). Dabei ist die Tatsache, dass eine zunächst schwierig klingende Aufgabe tatsächlich leicht lösbar ist, einem Großteil der Teilnehmer im Vorhinein bekannt. Auch eine Abwandlung dieses Ansatzes, bei dem die Teilnehmer zunächst eine Trainingsaufgabe mit guten oder schlechten Ergebnislisten lösen, um auf diese Weise die Erwartungshaltung für die folgenden Aufgaben zu manipulieren wird hier nicht verfolgt. Zum einen ist nicht klar, ob eine einzelne Testaufgabe

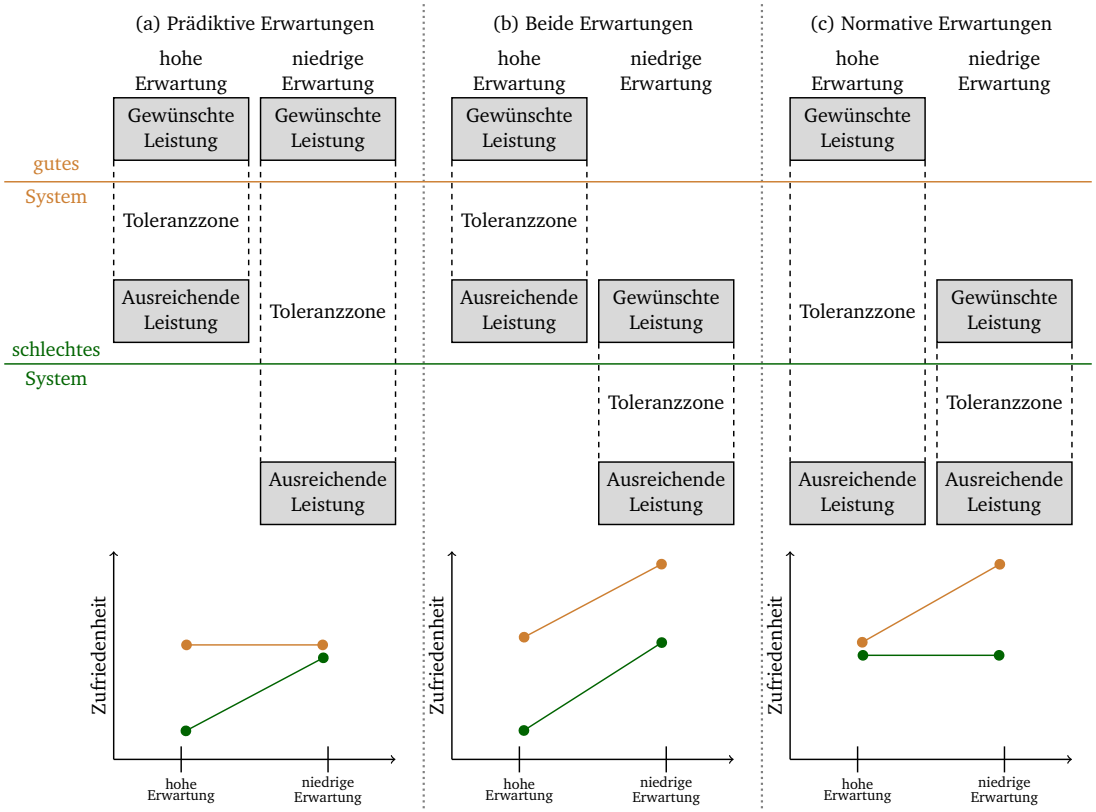


Abb. 4.3.: Schematische Darstellung möglicher Auswirkungen der Erwartungsmanipulation auf die Erwartungshaltung der Testteilnehmer (oben) und resultierende Zufriedenheitsreaktionen (unten). Die Bilder (a) und (c) stellen die Extremfälle dar, in denen die Manipulation allein die prädiktiven bzw. normativen Erwartungen verändert, während in Bild (b) beide Grenzen der Toleranzzone beeinflusst werden. Farbige horizontale Linien geben beispielhaft die wahrgenommene Leistung des guten bzw. schlechten Systems an. Im unteren Teil der Abbildung finden sich die nach dem C/D-Paradigma zu erwartenden Zufriedenheitsreaktionen. Hierbei resultiert eine wahrgenommene Systemleistung innerhalb der Toleranzzone in einer konstanten Zufriedenheit, während eine Leistung außerhalb der Toleranzzone zu einer positiven bzw. negativen Diskonfirmation führt.

ausreicht, um die entsprechende Erwartungshaltung hervorzurufen und zum anderen würde dies zu einem für die Testpersonen umfangreicheren Testablauf führen. Im Sinne der in Abschnitt 2.2.2 diskutierten Selbstwirksamkeitserwartung wäre in einem solchen Fall zudem denkbar, dass sich im Suchprozess Selbstwirksamkeits- und Systemerwartungen überlagern und somit die persönliche Wahrnehmung des eigenen Suchbeitrags individuelle Sucherfolgserwartungen hervorbringt. Stattdessen wird entschieden, eine Kombination der Ansätze von Szajna und Scamell (1993), Cuadra und Katter (1967) und Jansen et al. (2007) zu wählen, da in allen drei Arbeiten die Effektivität der jeweiligen Methode demonstriert werden kann. In diesem Sinne wird eine Manipulation der Erwartungshaltung auf Basis von schriftlichen bzw. auf Video festgehaltenen Testinstruktionen favorisiert. Die Art der Manipulation orientiert sich dabei an den Ergebnissen für die Markenwahrnehmung von Jansen et al. (ebd.). Das konkrete Vorgehen wird jedoch im Verlauf der drei Experimente variiert und verfeinert. Der Grund dafür liegt in der Schwierigkeit der Abgrenzung des beobachteten Erwartungsunterschieds von zufälligen Schwankungen. Wie in Abschnitt 2.1.1.1 bereits angedeutet, gilt es, eine Manipulationsmethode zu finden, die diesen

Unterschied so einstellt, dass die Toleranzzonen der beiden experimentellen Gruppen sich möglichst nicht überschneiden, um einen Einfluss der Erwartungshaltung im Rahmen der Studien tatsächlich sichtbar zu machen (Zeithaml et al., 1993). Dies betrifft zum einen die inhaltliche Präsentation der beiden Erwartungshaltungen, für die drei unterschiedliche Ansätze gewählt werden: Den Testpersonen wird mitgeteilt, dass sie mit einem professionellen oder einem von Studenten entwickelten System arbeiten (Experiment 1). Den Teilnehmern werden zwei Systeme präsentiert, die jeweils als gutes bzw. schlechtes System vorgestellt werden (Experiment 2). Den Versuchspersonen wird mitgeteilt, dass die Studie zwei Systeme vergleichen möchte und sie dem besseren bzw. schlechteren System zugeteilt sind (Experiment 3). Zum anderen wird auch die Instruktionsweise variiert: Angefangen von einem schriftlichen Instruktionstext, über ein Einführungsvideo in dem der Testleiter zu sehen ist, hin zu einer Kombination aus schriftlicher und akustischer Darbietung, bei welcher der Instruktionstext in Form einer Powerpoint-Präsentation mit einem Audiofile hinterlegt ist. Das detaillierte Vorgehen kann dabei in den entsprechenden Abschnitten 5.3.1, 6.3.1 und 7.3.1 der einzelnen Experimente nachvollzogen werden.

Obgleich die hier dargestellte Manipulationsstrategie einen plausiblen Ansatz zur Änderung der Erwartungshaltung der Testpersonen darstellt, kann aufgrund der fehlenden Erfahrungen mit Erwartungen im Kontext der Informationssuche noch keine abschließende Hypothese in Bezug auf die genaue Wirkungsweise dieses Ansatzes formuliert werden. Insbesondere bleibt die Frage offen, ob das gewählte Vorgehen in stärkerem Maße die prädiktiven oder normativen Erwartungen der Testpersonen beeinflusst. In Bezug auf die in Abschnitt 2.1.1.2 diskutierten Determinanten der Erwartungsbildung wäre prinzipiell eine Änderung beider Erwartungskomponenten möglich, da die dargebotenen Einführungstexte sowohl als indirekte Kommunikation über die Suchleistung als auch als Hinweis auf die zu erwartende Leistung gewertet werde könnte. Während sich nach Zeithaml et al. (ebd.) im ersten Fall die normativen Erwartungen ändern würden (so soll es sein), wäre im zweiten Fall ein stärkerer Einfluss auf die prädiktiven Erwartungen (so wird es sein) anzunehmen (vgl. Abb. 2.2). In Bezug auf den in der Kundenzufriedenheitsforschung postulierten Toleranzbereich, innerhalb dessen sich beim Nutzer Zufriedenheit einstellt, beeinflussen normative Erwartungen jedoch gerade die obere, prädiktive Erwartungen hingegen die untere Grenze dieser Akzeptanzzone.

Abbildung 4.3 zeigt exemplarisch, welche Konsequenzen diese beiden denkbaren Auswirkungen der Manipulationsstrategie auf das Verhältnis der Toleranzbereiche bei hoher und niedriger Erwartungshaltung haben könnten. Unter der Maßgabe des vom C/D-Paradigmas angenommenen Soll-Ist-Vergleichs für die Zufriedenheitsbildung führen beide Fälle zu unterschiedlichen Vorhersagen für die Zufriedenheitsreaktion der Probanden, womit in gewisser Weise die tatsächlich auftretende Beeinflussung der Erwartungshaltung in den Nutzerstudien mit überprüft werden kann. Des Weiteren stellt Abbildung 4.3 auch den Fall einer Beeinflussung beider Erwartungskomponenten dar (Bild (b)). Allerdings ist zu beachten, dass Zeithaml et al. (ebd.) als weitere Determinante der Erwartungsbildung auch die Wahrnehmung der Selbstwirksamkeit nennt, die, wie bereits angedeutet, gerade im Kontext der Informationssuche die Erwartungsbildung zusätzlich beeinflussen könnte.

Dies schließt die Diskussion über das Vorgehen zur Operationalisierung und Manipulation der unabhängigen Variablen ab. In den folgenden Abschnitten liegt der Fokus auf der Erhebung der

abhängen Variablen Benutzerleistung und Benutzerzufriedenheit.

4.2.2. Abhängige Variablen der Benutzer-System-Interaktion

Nachdem der vorangegangene Abschnitt das Stimulusmaterial beschreibt, stellt der nun folgende Abschnitt die Verfahren vor, mit denen die abhängigen Variablen erfasst werden können. Wie in Kapitel 3 diskutiert, sind die abhängigen Variablen dieser Arbeit der objektiv erreichte Sucherfolg der einzelnen Teilnehmer sowie die von ihnen subjektiv empfundene Zufriedenheit. Da sowohl die verfolgte Strategie zur Erfassung des individuellen Sucherfolgs als auch die erhobene Gesamtzufriedenheit (vgl. Kap. 3) in engem Zusammenhang mit der wahrgenommenen Relevanz der Suchergebnisse stehen, werden im Folgenden zunächst verschiedene Ansätze zur Messung der durch die Benutzer wahrgenommenen Relevanz betrachtet.

4.2.2.1. Verfahren zur Relevanzmessung

Der folgende Abschnitt bietet zunächst eine Übersicht über im Kontext von IIR-Studien verwendete Verfahren zur Erfassung der Relevanz von Dokumenten. Wie bereits in Abschnitt 3.1 im Rahmen der Zielfaktoren dieser Arbeit diskutiert, enthält der Prozess der Relevanzbeurteilung subjektive Komponenten, die schwer objektiviert oder nachprüfbar gemacht werden können. Dennoch bildet die Messung von Relevanz nach wie vor die entscheidende Basis für die Evaluierung von IR-Systemen. Forschungsarbeiten, die sich gezielt mit diesem Thema auseinandersetzen, lassen sich grob in zwei Gruppen unterteilen. Ein Ansatz besteht in Anlehnung an traditionelle Bewertungsmaße wie Recall und Precision in der Verwendung binärer bzw. kategorialer Relevanzurteile. Allerdings wird an diesem Ansatz kritisiert, dass sich das vielschichtige Bewertungsverhalten von Benutzern nur unzureichend durch kategoriale Skalen abbilden lasse. Aus diesem Grund greifen Studien der zweiten Gruppe auf kontinuierliche Bewertungskalen zurück. Im Folgenden werden Untersuchungen aus beiden Bereichen vorgestellt.

Die große Mehrheit der experimentellen Studien zum Informationssuchverhalten können der erstgenannten Gruppe zugeordnet werden (Cuadra u. Katter, 1967; Saracevic et al., 1988; Purgailis Parker u. Johnson, 1990; Smithson, 1994; Tombros u. Sanderson, 1998; Belkin et al., 1999; Hersh et al., 2000; Huang u. Wang, 2004; Allan et al., 2005; Al-Maskari et al., 2006; Kelly et al., 2007; Smith u. Kantor, 2008; Smucker u. Jethani, 2010a). Dies ist auf verschiedene Faktoren zurückzuführen, wie z.B. die lange Tradition binär-orientierter Relevanzurteile im IR oder den Wunsch auf gut etablierte Evaluierungsmaße zurückgreifen zu können. Auch beinhalten die meisten Testkollektionen (wie TREC u. CLEF) in der Regel binäre oder zumindest kategoriale Relevanzurteile. Allerdings gibt es zum jetzigen Zeitpunkt in der Literatur noch keinen Konsens hinsichtlich der *richtigen* Skalierungsweise respektive der optimalen Anzahl von Skalenstufen. Viele Autoren gehen davon aus, dass Relevanz typischerweise binär ist und beschränken die Messung somit auf eine dichotome Skala (Tombros u. Sanderson, 1998; Belkin et al., 1999; Hersh et al., 2000; Allan et al., 2005; Al-Maskari et al., 2006; Kelly et al., 2007; Smucker u. Jethani, 2010a). Eine zum Teil vorgenommene Erweiterung besteht in dem Hinzufügen einer Mittelkategorie, was den Vorteil bietet, dass unentschiedene Meinungen (teilweise relevant) ausgedrückt werden können (Saracevic et al., 1988; Purgailis Parker u. Johnson, 1990; Smith u. Kantor, 2008). Spink und Greisdorf (2001) halten die Vorgabe einer solchen Ausweichkategorie für unbedingt erforderlich, da diese Art der Skalierung eine eindeutigere Positionierung der Befrag-

ten ermöglicht. In ihrer Studie zum Vergleich unterschiedlicher Skalierungsverfahren steht den Befragten eine 4-stufige Bewertungsskala mit den Ausprägungen *relevant*, *teilweise relevant*, *teilweise irrelevant* sowie *irrelevant* zur Verfügung. In absoluten Zahlen entfallen 376 der insgesamt abgegebenen 1059 Relevanzurteile auf den mittleren Bereich der Skala (Spink u. Greisdorf, 2001, S. 165). Aufgrund dieses mit deutlich über 30% recht hohen Anteils unentschiedener Urteile stellen die Autoren die Sinnhaftigkeit von auf binären Relevanzurteilen basierenden Precisionmaßen in Frage und empfehlen stattdessen die Verwendung eines von Greisdorf und Spink (2001) eingeführten Median-Maßes, um mögliche Verzerrungen durch unentschiedene Meinungen zu vermeiden. In ähnlicher Weise lässt Sormunen (2002) eine Neubewertung von 38 TREC-Topics durchführen. Anstelle der ursprünglich von TREC verwendeten binären Relevanzskala steht den Juroren wie bei Greisdorf und Spink (2001) eine 4-stufige Skala zur Verfügung (Sormunen, 2002, S. 325). Im Vergleich mit den binären Relevanzurteilen stellt sich heraus, dass die Hälfte aller als relevant bewerteten Dokumente auf der erweiterten Skala nur als geringfügig relevant eingestuft werden (ebd., S. 326 f.).

Gelegentlich werden auch Skalen mit mehr Abstufungen gewählt. Smithson (1994) schlagen sechs, Eisenberg und Barry (1988) und Huang und Wang (2004) sieben und Cuadra und Katter (1967) neun Abstufungen vor. Zwar bietet eine hohe Anzahl von Abstufungen den Vorteil einer sehr differenzierten Messung und Analyse von Relevanz. Es ist jedoch laut Küchenhoff (2006, S. 165) darauf zu achten, dass die Differenzierungsfähigkeit der Befragten, zwischen den einzelnen Abstufungen zu unterscheiden, nicht überschätzt wird: „In der Praxis haben sich Rating-Skalen mit fünf bis sieben Ausprägungen bewährt." Tang et al. (1999, S. 261) vergleichen Skalen mit zwei bis elf Abstufungen und zeigen, dass die Beurteilungssicherheit der Testpersonen zunimmt je mehr Abstufungen eine Relevanzskala aufweist. Auf Basis ihrer Untersuchungsergebnisse empfehlen sie die Verwendung einer 7-stufigen Skala mit oder einer 6-stufigen Skala ohne neutrale Mittelkategorie (ebd., S. 263). Dies wird auch durch Moosbrugger und Kelava (2011, S. 51) gestützt, die allgemein in Bezug auf die Fragebogenkonstruktion davon ausgehen, dass mehr als sieben Antwortkategorien nicht zu einem Informationsgewinn beitragen.

Wie bereits angedeutet, besteht ein zweiter Ansatz darin Relevanz nicht kategorial, sondern auf einer kontinuierlichen Skala zu erfassen. Studien, die diesen Ansatz verfolgen, gehen davon aus, dass es sich bei Relevanz um eine stetige Größe handelt und ein solches Vorgehen weniger anfällig für Verzerrungen, wie z.B. Antworttendenzen oder Reihenfolgeeffekte, ist (vgl. Abschn. 4.2.3.2). Methodisch wird dabei häufig auf die von Stevens (1975) im Kontext der Wahrnehmungspsychologie entwickelte und auch im Rahmen der Einstellungsforschung verwendete *Magnitude-Estimation-Methode* zurückgegriffen (Eisenberg u. Hu, 1987; Eisenberg u. Barry, 1988; Eisenberg, 1988; Janes, 1991b; Janes, 1991a; Bruce, 1994; Spink u. Greisdorf, 2001; Greisdorf u. Spink, 2001). Dabei werden die Testteilnehmer aufgefordert, die Stärke einer Wahrnehmung, hier die Relevanz eines Dokuments, durch die Nennung eines Zahlenwerts, das Setzen eines Kreuzes auf einer Skala ohne konkrete Skalenstufen oder die Länge einer Linie zu bewerten (Bortz u. Döring, 2006). Janes (1991a, S. 432) sieht einen der Vorteile dieser Methode in dem größeren Angebot an statistischen Auswertungsmöglichkeiten: „The chief advantage of using magnitude estimation techniques is that they provide the researcher with ratio-level data, which can be used in more sophisticated statistical techniques than ordinal-level data, obtained from

categorical relevance judgments (relevant/ partially relevant/ not relevant, etc.)." In seiner Studie vergleicht Janes (ebd.) den Einfluss verschiedener Metadaten auf die Relevanzbeurteilung von 40 Testpersonen. Die Bewertung erfolgt mit Hilfe einer zehn Zentimeter langen Magnitude-Estimation-Skala. Als problematisch stellt sich heraus, dass fast ein Viertel der Testpersonen die Bewertungsskala zu schnell ausschöpft und dadurch gerade am oberen und unteren Ende der Relevanzskala keine genauere Differenzierung mehr vorgenommen werden kann (ebd., S. 639). Ein Effekt, der auch als *Decken-* und *Boden-Effekt* bezeichnet wird. Auch Bruce (1994) verwendet diese Methode dazu, die relative Wichtigkeit verschiedener Dokumenteigenschaften für die Relevanzurteile der Benutzer zu bestimmen (vgl. Abschn. 3.1.3). Die im Folgenden dargestellten Studien hingegen vergleichen beide Ansätze in Bezug auf ihre Eignung zur Relevanzerfassung.

Die Studien von Eisenberg und Hu (1987) und Janes (1991b) bspw. untersuchen inwieweit die Ergebnisse kontinuierlicher Relevanzeinschätzugen einen binärer Character von Relevanzurteilen rechtfertigen können. Dazu bitten sie Testpersonen Relevanzbeurteilungen mit Hilfe einer Magnitude-Estimation-Skala vorzunehmen. Beide Studien können die Existenz eines eindeutigen Schwellenwerts, der relevante von irrelevanten Dokumenten trennt, nicht belegen. Greisdorf und Spink (2001) und Spink und Greisdorf (2001) vergleichen vier Verfahren zur Skalierung von Relevanzurteilen. Die getesteten Verfahren umfassen jeweils eine Magnitude-Estimation-Skala, eine 4-stufige Kategorienskala mit Ausweichkategorien, fünf dichotome Relevanzvariablen (Format, Inhalt, Angemessenheit, Nutzen u. Motivation) sowie eine offen gestellte Frage ohne Vorgabe von Antwortmöglichkeiten. Beide Untersuchungen stellen fest, dass Nutzer in der Lage sind mit einer Vielzahl von Skalen umzugehen: „Any scale that implies some range of utility, value, strength, importance or magnitude appears to suit the end-user in making a relevance assessment of items retrieved from an IR system." (Greisdorf u. Spink, 2001, S. 853) Darüber hinaus können Greisdorf und Spink (ebd., S. 853) zeigen, dass die getesteten Skalen alle zu einer bimodalen Häufigkeitsverteilung führen, was als Hinweis darauf gedeutet werden kann, dass eine dichotome Antwortskala im Prinzip ausreichend ist.

Vor diesem Hintergrund wird im Rahmen dieser Arbeit zu Gunsten der Kompatibilität mit etablierten System- und Benutzerleistungsmaßen sowie der Vergleichbarkeit mit anderen Studien auf die Verwendung von Magnitude-Estimation-Methoden verzichtet. Stattdessen bewerten die Teilnehmer in den ersten beiden Experimenten die Relevanz der Dokumente wie im überwiegenden Teil der Literatur anhand einer binären Skala als relevant bzw. irrelevant. Im dritten Experiment hingegen wird auf eine erweiterte kategoriale Skala mit acht Bewertungskategorien zurückgegriffen. Dies entspricht nahezu der von Tang et al. (1999) und Moosbrugger und Kelava (2011) empfohlenen Größenordnung von sieben Antwortabstufungen, erlaubt andererseits jedoch die Umrechnung in eine binäre Skala indem jeweils die ersten bzw. letzten vier Kategorien zusammengefasst werden. Darüber hinaus ermöglicht die Achtstufigkeit der Skala auch einen direkten Vergleich mit den im dritten Experiment verwendeten 4-stufigen Jurorenurteilen (vgl. Abschn. 4.1.3.3) durch Kombination von jeweils zwei aufeinander folgenden Antwortkategorien. Nach dieser Diskussion unterschiedlicher Relevanzskalen, widmet sich der folgende Abschnitt konkreten Verfahren zur Sucherfolgsmessung, die auf diesen Relevanzbewertungen der Nutzer aufbauen.

Tab. 4.3.: Übersicht über in der Literatur verwendete Benutzerleistungsmaße.

Maß	Beschreibung	Verwendet in
Anz. rel. bew. Dok.	Anzahl der Dokumente, die als relevant bewertet werden.	Al-Maskari und Sanderson (2010), Veerasamy und Belkin (1996)
Anz. richtig rel. bew. Dok.	Anzahl der Dokumente, die in Übereinstimmung mit den Juroren als relevant bewertet werden.	Turpin und Scholer (2006), Al-Maskari et al. (2008b), Smucker und Jethani (2010a), Al-Maskari und Sanderson (2010)
Instance Recall	Anzahl gefundener Dokumente die unterschiedliche inhaltliche Facetten des Suchthemas enthalten.	Allan et al. (2005), Turpin und Hersh (2001)
Benutzerrecall (BR)	Anteil der vom Nutzer richtig als relevant bewerteten Dokumente an allen relevanten Dokumenten im Testkorpus.	Al-Maskari et al. (2006), Käki und Aula (2008)
Benutzerprecision (BP)	Anteil der vom Nutzer richtig als relevant bewerteten Dokumente an allen vom Nutzer als relevant bewerteten Dokumenten.	Al-Maskari et al. (2006), Käki und Aula (2008)
Interaktive TREC Precision	Anteil aufgerufener relevanter Dokumente an allen aufgerufenen Dokumenten.	Veerasamy und Belkin (1996)
Interaktive Benutzerprecision	Anteil als relevant bewerteter Dokumente an allen aufgerufenen Dokumenten.	Veerasamy und Belkin (1996)
Nutzerselektivität	Anteil der nicht aufgerufenen Dokumente.	Smith und Kantor (2008)

4.2.2.2. Verfahren zur Sucherfolgsmessung

Ähnlich wie im Fall des systemzentrierten Ansatzes (vgl. Abschn. 4.2.1.1), stellt sich auch im Kontext der benutzerorientierten Evaluierung die Frage, wie der individuelle Sucherfolg bzw. die Benutzerleistung erfasst werden kann. Prinzipiell lassen sich zwei Ansätze zur quantitativen Bestimmung der Benutzerleistung unterscheiden, je nachdem, ob die Leistung aus der Perspektive des Systems oder des Benutzers erhoben wird. Während im ersten Fall an den systemorientierten Ansatz angelehnte Maße zum Einsatz kommen, wird im zweiten Fall der Effizienz der Suche aus Benutzersicht ein stärkeres Gewicht beigemessen. Darüber hinaus verfolgen einige Studien den Ansatz, den wahrgenommenen Sucherfolg direkt beim Nutzer zu erfragen. Da dieser Aspekt jedoch eng mit der Nutzerzufriedenheit verbunden ist, werden solche Ansätze im Kontext der Zufriedenheitsmessung in Abschnitt 4.2.2.3 erläutert. Da viele Aspekte, insbesondere auch in Bezug auf den Stand der Forschung bereits in Abschnitt 3.2 im Zusammenhang mit der Bewertung des individuellen Sucherfolgs behandelt sind, gibt der folgende Abschnitt nur einen kurzen Überblick über die in der Literatur verwendeten objektiven Maße, um daran anschließend die für die vorliegende Arbeit gewählte Strategie zur Erfassung der Benutzerleistung darzustellen.

Die Mehrheit der in der Literatur verwendeten Benutzerleistungsmaße beruht auf entsprechenden Vorbildern aus dem systemorientierten Evaluierungskontext. Dabei wird die Rolle der von einem Suchsystem zurückgelieferten Ergebnislisten nun von Relevanzbewertungen der Testeilnehmer übernommen. Im Wesentlichen beruhen alle diese Maße auf einem Vergleich zwischen einem als objektiv betrachteten Jurorenurteil und der Relevanzbewertung der Testteilnehmer. Dies führt wie im systemzentrierten Fall dazu, dass zur Durchführung einer solchen Nutzerstudie auf ein zuvor bewertetes Testkorpus zurückgegriffen werden muss. Eine Übersicht über in der

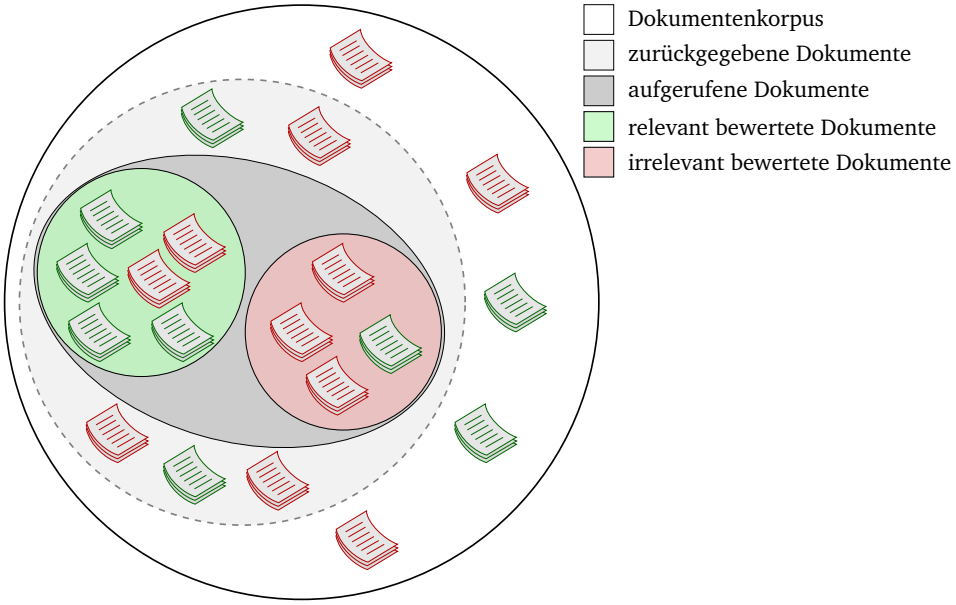


Abb. 4.4.: Schematische Darstellung einzelner Dokumentenmengen zur Beurteilung der Benutzerleistung. Markiert sind neben dem Dokumentenkorpus und der Menge der von der Suchmaschine zurückgelieferten Dokumente auch die vom Nutzer als relevant bzw. irrelevant bewerteten Dokumente. Von den Juroren als relevant annotierte Dokumente sind grün, als irrelevant bewertete Dokumente hingegen rot hinterlegt.

Literatur verwendete Verallgemeinerungen klassischer Systemleistungsmaße auf den Nutzerkontext ist in Tabelle 4.3 zusammengefasst. Der Benutzerrecall ist hier bspw. als Verhältnis zwischen der Anzahl relevanter Dokumente im Korpus und der Anzahl vom Nutzer als relevant erkannten Dokumente definiert.

Allgemeiner können die in Tabelle 4.3 aufgelisteten Leistungsmaße als Größe bzw. Verhältnis spezieller Untermengen der Testkollektion verstanden werden. Dieser Ansatz ist graphisch in Abbildung 4.4 dargestellt: Ausgehend von der Menge aller im Testkorpus vorhandenen Informationsobjekte unterscheidet man zunächst die von den Juroren als relevant bzw. irrelevant annotierten Dokumente. Stellt ein Nutzer während einer Suchsession mehrere Suchanfragen, werden die Einträge in den zurückgegebenen Ergebnislisten in der Menge der zurückgelieferten Dokumente zusammengefasst, die sowohl irrelevante, als auch relevante Quellen enthält. Aus diesem Pool wählt die Testperson nun Informationsobjekte zum Betrachten aus und fällt ihr subjektives Relevanzurteil. Daraus ergeben sich die Mengen der durch die Testperson als relevant bzw. irrelevant bewerteten Dokumente. Aufbauend auf diesen Untermengen lassen sich sowohl die Benutzerleistungsmaße aus Tabelle 4.3 als Verhältnisse darstellen, als auch weitere Benutzerleistungsmaße definieren. Dies erlaubt bspw. neben der Übereinstimmungstendenz zwischen Juroren- und Probandenurteil in Bezug auf relevante Dokumente, wie sie durch die Benutzerprecision beschrieben wird, auch Vergleiche hinsichtlich der als irrelevant bewerteten Quellen. Darüber hinaus werden im Rahmen dieser Arbeit insbesondere auch als *Imprecisionmaße* bezeichnete Verhältnisse betrachtet, die anstelle der Übereinstimmungs- die Widerspruchstendenz zwischen Juroren und Testteilnehmern quantifizieren. So ergibt sich bspw. eine Art Benutzerimprecision als Anteil der vom Nutzer im Widerspruch zu den Juroren als relevant bewerteten Dokumente an allen vom Nutzer als relevant bewerteten Quellen.

Um für die hier durchgeführten Nutzerstudien ein möglichst breites Spektrum an Benutzerleistungsfacetten abfragen zu können, wird in den folgenden Auswertungskapiteln der Einfluss von Systemleistung und Benutzererwartung auf eine ganze Reihe derart definierter Benutzerleistungsmaße untersucht. Die jeweils speziell gewählten Maße können den entsprechenden Abschnitten 6.3.2 und 7.3.2 entnommen werden, welche die für das jeweilige Experiment gewählte Operationalisierung der Benutzerleistung darstellen. Allerdings sei an dieser Stelle kurz angemerkt, dass sich das Aufteilen in entsprechende Untermengen in natürlicher Weise auf nichtbinäre Relevanzurteile sowohl in Bezug auf das zugrundeliegende Testkorpus als auch in Bezug auf die Benutzerurteile übertragen lässt (vgl. Abschn. 7.3.2). Die Methode erlaubt also auch die Verwendung feinerer Relevanzskalen, die einen detaillierteren Blick auf bspw. die dynamische Entwicklung der Relevanzwahrnehmung der Probanden zulassen.

Tab. 4.4.: Übersicht über in der Literatur verwendete Effizienz- und Aufwandsmaße.

Maß	**Beschreibung**	**Verwendet in**
Mittlere Rankingposition	Mittelwert über die Rankingpositionen aller Dokumente, die vom Nutzer als relevant bewertet werden	Toms et al. (2005)
Maximale reziproke Rankingposition	Kehrwert der maximalen Rankingposition gemittelt über alle Suchanfragen	Radlinski et al. (2008)
Mittlere reziproke Rankingposition	Summe der Kehrwerte der Rankingpositionen aller betrachteten Dokumente gemittelt über alle Suchanfragen	Radlinski et al. (2008)
RankFRD	Rankingposition des ersten richtig relevant bewerteten Dokuments	Al-Maskari und Sanderson (2010)
TimeFRD	Zeit zum Auffinden des ersten richtig relevanten Dokuments	Al-Maskari et al. (2008b), Turpin und Scholer (2006), Scholer und Turpin (2009), Al-Maskari und Sanderson (2010)
Search Speed	Gefundene relevante Dokumente pro Minute	Käki und Aula (2008)
Listenverweildauer	Zeit, die zum Lesen der Suchergebnisliste benötigt wird	Toms et al. (2005)
Dauer der Suche	Zeit die zur Bearbeitung der Suchaufgabe benötigt wird	Allan et al. (2005), Smith und Kantor (2008), Xu und Mease (2009)
Suchanfragenrate	Anzahl der Suchanfragen, die pro Minute eingegeben werden	Smith und Kantor (2008)
Anzahl Suchanfragen	Anzahl Suchanfragen pro Session/ Aufgabe	Järvelin (2009), Toms et al. (2005), Radlinski et al. (2008), Al-Maskari et al. (2008b), Al-Maskari und Sanderson (2010)
Reformulierungsrate	Anteil der Suchanfragen auf die innerhalb einer Suchsession eine weitere Suchanfrage folgt	Radlinski et al. (2008)
Abbruchrate	Anteil der Suchanfragen für die kein Dokument betrachtet wird	Radlinski et al. (2008)
Immediate Accuracy	Anteil der Suchaufgaben bei denen ein Nutzer nach dem Öffnen von n Dokumenten mindestens ein relevantes Dokument gefunden hat	Käki und Aula (2008)

In Ergänzung zu diesen auf dem systemzentrierten Evaluierungsansatz aufbauenden Kennzahlen finden in der Literatur auch Benutzerleistungsmaße Anwendung, die sich enger an der

Interaktion zwischen Nutzer und System und ihrer Effizienz orientieren und ebenfalls nur im Rahmen von Benutzertests zugänglich sind. Dies ist insbesondere vor dem Hintergrund zu sehen, dass die in Abschnitt 3.2.1 dargestellten Studien häufig Kompensationseffekte in Bezug auf bspw. den Benutzerrecall beobachten und somit eine Notwendigkeit alternativer Messverfahren besteht. Andererseits stellen Lancaster und Warner (1993) in Bezug auf die Benutzerprecision fest, dass sie als ein indirektes Maß für die investierte Zeit und den Aufwand der Nutzer interpretiert werden kann, da bei hoher Precision Zeit und Aufwand zum Auffinden relevanter Dokumente geringer ausfallen. Lancaster und Warner (ebd.) plädieren deshalb für die Verwendung von Zeitmaßen. Tabelle 4.4 gibt einen Überblick über solche in der Literatur gebräuchliche Benutzereffizienzmaße. Der eine Teil dieser Maße bezieht dabei neben den beschriebenen Dokumentenmengen auch explizit die Rankingposition der aufgerufenen Dokumente mit in die Betrachtung ein, um den Aufwand zu quantifizieren. Bei dem anderen Teil rückt hingegen der zeitliche Aufwand bzw. die Zahl der Interaktionen, die Nutzer im Rahmen einer Suchsession durchführen müssen, in den Fokus. Insbesondere werden also auch Kennzahlen in die Analyse einbezogen, die über die einzelne Suchanfrage hinaus gehen, und den gesamten Suchprozess betrachten. Die Ergebnisse einer Studie von Järvelin (2009), wonach Anfragen, die in der Einzelevaluation wenig Sinn ergeben im Kontext einer gesamten Suchsession dennoch zu guten Ergebnissen führen, zeigen, dass dieses Vorgehen sinnvoll ist. Im Sinne einer möglichst umfassenden Analyse von System- und Erwartungseinfluss auf das Nutzerverhalten wird deshalb entschieden, auch in den im Rahmen dieser Arbeit durchgeführten Experimenten eine Reihe von Zeit- und Effizienzmaßen zu erheben, die im einzelnen den Abschnitten 6.3.2 und 7.3.2 entnommen werden können.

Zusammenfassend enthält dieser Abschnitt einen Überblick über die im Rahmen interaktiver IR-Studien gebräuchlichsten Benutzerleistungs- und Effizienzmaße, auf die auch im Rahmen dieser Arbeit zurückgegriffen werden kann. Die Darstellung der Leistungsmaße als Verhältnisse von Untermengen der vom Nutzer aufgerufenen Dokumente, ermöglicht darüber hinaus die Definition weitergehender Kennzahlen, die es bspw. erlauben die Widerspruchstendenz zwischen Benutzer- und Jurorenurteilen zu quantifizieren. Auch vor dem Hintergrund der teilweise widersprüchlichen Studienlage zur Rolle der Systemleistung im Benutzerkontext, wie sie in Abschnitt 3.2.1 dargestellt ist, scheint es sinnvoll, eine möglichst umfassende Menge von Leistungs- und Effizienzmaßen in die Untersuchung einzubeziehen. Bisher nicht berücksichtigt sind in diesem Abschnitt Ansätze, die Leistung aus der Benutzerperspektive zu quantifizieren. Da dieses Vorgehen jedoch eng mit der Zufriedenheit der Nutzer mit der eigenen Suchleistung bzw. der Systemleistung verbunden ist, wird dieser Aspekt im nun folgenden Abschnitt im Zusammenhang mit der Zufriedenheitsmessung diskutiert.

4.2.2.3. Verfahren zur Zufriedenheitsmessung

In diesem Abschnitt werden Verfahren und Fragebogeninstrumente zur Erfassung der Nutzerzufriedenheit vorgestellt und diskutiert. Im Gegensatz zur Benutzerleistung handelt es sich hierbei um eine latente Variable, die sich als subjektive Wahrnehmung des Benutzers einer direkten Beobachtung entzieht (Bortz und Döring, 2006, S. 4; Kelly, 2009, S. 37). Um im Rahmen experimenteller Studien zum Informationssuchverhalten trotzdem Aufschluss über die Nutzerzufriedenheit zu erlangen, ist es deshalb notwendig, geeignete Indikatoren zu finden, die einer direkten Messung zugänglich sind. Im Falle der Zufriedenheit geschieht dies typischerweise über Items in

einem Fragebogen, die unterschiedliche Dimensionen des Konstrukts Nutzerzufriedenheit erfassen sollen (vgl. Abschn. 3.3.2). Obwohl die Nutzerzufriedenheit Gegenstand unterschiedlichster IIR-Studien ist, steht zu diesem Zeitpunkt noch kein standardisiertes und evaluiertes Instrument zur Messung der Zufriedenheit im Suchmaschinenkontext zur Verfügung. Kelly (2009, S. 37) bemerkt dazu „Unfortunately, there are not a lot of research programs focused on measurement, which makes it difficult to understand the extent of measurement problems in IIR evaluations. Many measures are developed in an *ad-hoc* fashion and there are few well-established measures and instruments, especially for indirect observables."

In der Tat führt dies zu einem breiten aber schwer vergleichbaren Methodenspektrum bei der Erfassung der Nutzerzufriedenheit im IR-Kontext aus dem im Folgenden einige Studien herausgegriffen werden. Häufig wird die Zufriedenheit über zusätzliche Fragebögen im Anschluss an die Bearbeitung der Testaufgaben ermittelt. Dabei variieren Art und Form der verwendeten Frageitems stark zwischen den verschiedenen Untersuchungen. Während manche Studien vollständig auf direkte Zufriedenheitsstatements verzichten und stattdessen die Relevanzurteile (Jansen et al., 2007), die Dokumentenverweildauer (Hu et al., 2011) oder die Bewertung der Systemleistung (Kelly et al., 2007) zur Beurteilung der Zufriedenheit heranziehen, fragen Huffman und Hochster (2007) bspw. allgemein nach der Zufriedenheit, wohingegen Al-Maskari et al. (2007) spezieller die Zufriedenheit mit Genauigkeit, Abdeckung und Ranking der Retrievalergebnisse bewerten lassen. In einer stärker methodisch ausgerichteten Untersuchung hingegen analysiert Su (1998) 20 unterschiedliche Zufriedenheitsindikatoren im Hinblick auf ihre Korrelation mit dem von den Nutzern wahrgenommenen Gesamtsucherfolg. Sie kommt zu dem Schluss, dass sich die auf einer 7-stufigen Likert-Skala eingeschätzte Nützlichkeit der Suchergebnisse (*value of search results as a whole*) am besten als Indikator für den Gesamtsucherfolg eignet. Befunde von Tagliacozzo (1977) hinsichtlich inkonsistenter Nutzerbewertungen in Bezug auf die Aspekte *helpfulness* und *usefulness* unterstreichen darüber hinaus die Mehrdimensionalität des Konstrukts Nutzerzufriedenheit und die damit einhergehende Notwendigkeit diese unterschiedlichen Dimensionen im Rahmen der Fragebogenkonstruktion zu berücksichtigen. Auf die Problematik, dass die Praxis Nutzerzufriedenheit einzig über ein einzelnes zusammenfassendes Maß zu erfassen, häufig zu einer Überschätzung der Nutzerzufriedenheit führt wird von Applegate (1993) hingewiesen. Vor dem Hintergrund dieser Ergebnisse und dem Fehlen eines etablierten Fragebogeninstruments zur Erfassung von Benutzerzufriedenheit im IIR-Kontext, wird in der vorliegenden Arbeit der Ansatz verfolgt, nicht nur in der IIR-Forschung angewendete Methoden zu berücksichtigen, sondern auch auf gut evaluierte Fragebögen aus angrenzenden Disziplinen, wie der Informationssystemforschung zurückzugreifen.

Wie die bisherigen Ausführungen zu den Einflussfaktoren bei der Informationssuche vermuten lassen, gibt es keine einfache Definition oder ein allgemein anwendbares Maß für Zufriedenheit. Zwar wurden in der Informationssystemforschung in den vergangenen 30 Jahren unterschiedliche Verfahren zur Messung der Nutzerzufriedenheit entwickelt (Bailey u. Pearson, 1983; Ives et al., 1983; Doll u. Torkzadeh, 1988; Parasuraman et al., 1988; Omar u. Lascu, 1993), allerdings hat keines dieser Instrumente Eingang in die IR-Evaluierung gefunden. Dies könnte darin begründet liegen, dass diese Instrumente einerseits zu wenige Details bezüglich unterschiedlicher Dimensionen der Systemleistung und andererseits für den IR-Kontext irrelevante Faktoren wie

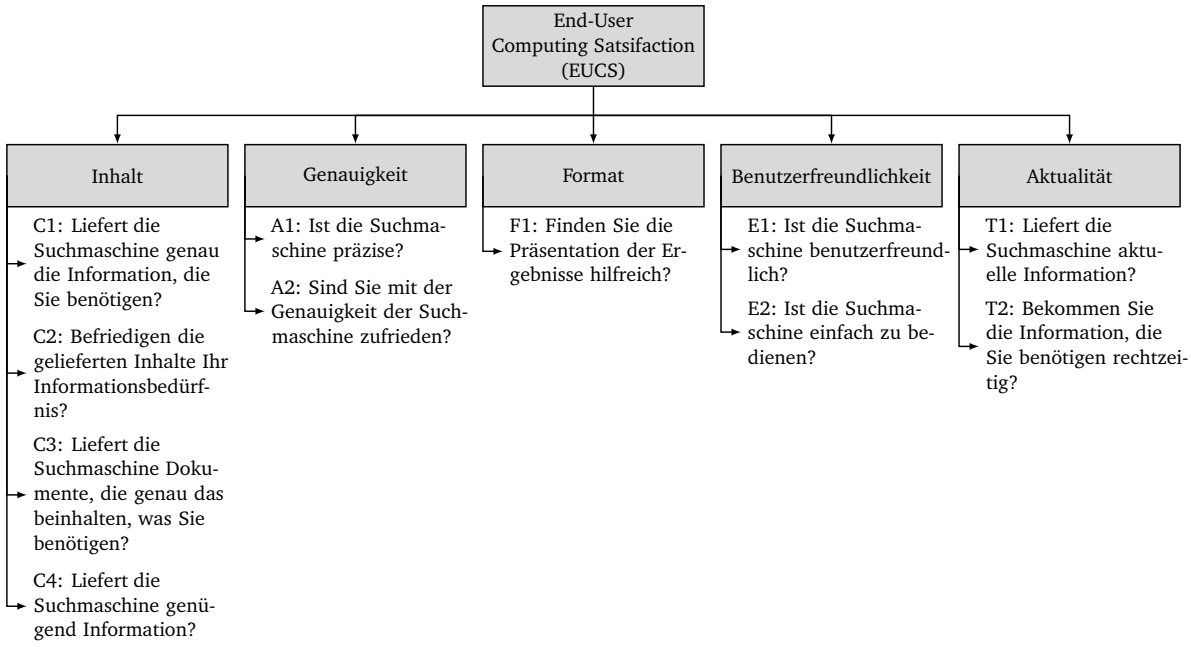


Abb. 4.5.: Die Zufriedenheitsfaktoren und Frageitems des EUCS-Instruments (nach Doll und Torkzadeh, 1988, S. 268). Das im Originalinstrument enthaltene Item F2: *Is the information clear?*, wird mangels einer präzisen Übersetzung, die eng am englischen Original bleibt, nicht berücksichtigt, um die Validität des Instruments nicht zu gefährden.

z.B. die Kompetenz des Kundendienstes erfassen. Im Folgenden werden einzelne dieser Ansätze vorgestellt. Ein von Bailey und Pearson (1983) entwickeltes Instrument umfasst 39 Faktoren. Als Skalierungsverfahren dient das *Semantische Differential* nach Osgood (1962), bei welchem jeder Faktor durch sechs 7-stufige Skalen erfasst wird, deren Extreme jeweils durch gegensätzliche Adjektivpaare beschrieben sind. Problematisch für den Einsatz im Rahmen der hier geplanten Untersuchungen erscheint vor allem die große Anzahl der Faktoren, deren Beantwortung bei der Bearbeitung mehrerer Suchaufgaben zu einer Ermüdung der Testteilnehmer führen könnte. Eine kürzere Fassung wird von Ives et al. (1983) vorgeschlagen und von Doll und Torkzadeh (1988) unter dem Namen *end-user computing satisfaction* (EUCS) weiterentwickelt, dessen Frageitems Abbildung 4.5 entnommen werden können. Dabei stellt dieses Instrument das derzeit wohl am häufigsten eingesetzte Fragebogenverfahren in der Informationssystemforschung dar. Neben entscheidungsunterstützenden Informationssystemen (McHaney et al., 1999) sind z.B. auch Systeme zur Unternehmensressourcenplanung (Somers et al., 2003) sowie Webportale (Xiao u. Dasgupta, 2002) mit diesem Verfahren erfolgreich evaluiert worden. Dabei umfasst das EUCS-Instrument die fünf in Abbildung 4.5 dargestellten, faktorenanalytisch ermittelten, Subskalen: Inhalt, Genauigkeit, Format, Gebrauchstauglichkeit und Aktualität (Doll et al., 1994; Doll u. Xia, 1997). Das Zufriedenheitsurteil wird mit einer 5-stufigen Likert-ähnlichen Skala erhoben. Inhaltlich scheint dieses Instrument für den IR-Kontext bereits gut geeignet, allerdings besteht insbesondere noch Erweiterungsbedarf in Bezug auf die Systemleistung, etwa zur Bewertung von Recall und Precision, aber auch in Bezug auf das Ranking der gefundenen Dokumente sowie den wahrgenommenen eigenen Sucherfolg. In einem weiteren von Omar und Lascu (1993) entwickelten Fragebogeninstrument bewerten die Testteilnehmer neben ihrer Zufriedenheit auch die Wichtigkeit der jeweiligen Items. Dieses Vorgehen erscheint auf der einen Seite sehr aufwändig,

auf der anderen Seite ergeben die Resultate von Omar und Lascu (1993) jedoch, dass derart gewichtete Zufriedenheitsurteile besser mit der Gesamtzufriedenheit korrelieren. Ein weiterer Vorteil dieser Methode ist in der Möglichkeit der genauen Identifikation der entsprechenden Determinanten der Zufriedenheitsreaktion zu sehen. Da aber nur einer der insgesamt fünf von Omar und Lascu (ebd.) herangezogenen Faktoren relevant für IR-Systeme ist, scheidet auch dieses Instrument für den Einsatz im IR-Kontext aus.

Anhand dieser Ausführungen wird deutlich, dass im Rahmen der vorliegenden Arbeit zwar auf vorhandene Instrumente zurückgegriffen und aufgebaut werden kann, sich jedoch keines der beschriebenen Instrumente direkt für den Einsatz im Kontext von experimentellen Studien zum Informationssuchverhalten eignet. Vor dem Hintergrund, dass es sich bei den EUCS-Items um ein bewährtes und validiertes Fragebogeninstrument handelt, wird in Experiment 2 und 3 ausgehend von diesen Items ein angepasster Fragebogen entwickelt. Die konkrete Umsetzung ist in den Abschnitten 6.3.2 und 7.3.2 der jeweiligen Kapitel erläutert. In den Abschnitten 6.4.4.2 und 7.4.4.2 hingegen finden sich Ergebnisse zur Replikation der für das EUCS-Instrument beschriebenen Faktorskalen mit Hilfe einer Hauptkomponentenanalyse (vgl. Abschn. 4.3.1). Ähnlich wie im Fall der Erwartungsmanipulation (vgl. Abschn. 4.2.1.2), wird das konkrete Vorgehen zur Erfassung der Benutzerzufriedenheit im Verlauf der drei Experimente variiert und verfeinert, um den Effekt der unabhängigen Variablen möglichst vollständig zu erfassen. Dies betrifft vor allem den Zeitpunkt der Zufriedenheitsmessung. Dieser ist im ersten Experiment so gewählt, dass die Probanden den Zufriedenheitsfragebogen nur einmal nach Bearbeitung aller Testaufaufgaben beantworten müssen. Jedoch kann mit diesem Vorgehen nur ein schwacher Einfluss von Systemleistung und Erwartungshaltung auf die Benutzerzufriedenheit beobachtet werden (vgl. Abschn. 5.4.3). Angesichts der Ergebnisse von Hu et al. (2011) (vgl. Abschn. 2.1.1.3) und Szajna und Scamell (1993) (vgl. Abschn. 3.3.2.4) ist jedoch zu vermuten, dass die Zufriedenheitswahrnehmung zeitlich begrenzt ist und unrealistische Erwartungen bereits nach der ersten Interaktion mit einem Suchsystem angepasst werden. Aus diesem Grund erfolgt die Zufriedenheitsmessung im Rahmen des zweiten und dritten Experiments direkt im Anschluss an die Bearbeitung der einzelnen Testaufgaben. Das detaillierte Vorgehen kann dabei in den entsprechenden Abschnitten 5.3.6, 6.3.6 und 7.3.6 der einzelnen Experimente nachvollzogen werden. Nachdem die Diskussion der Operationalisierung der abhängigen und unabhängigen Variablen im Rahmen der durchgeführten Experimente nun abgeschlossen ist, beschäftigt sich der folgende Abschnitt mit der Kontrolle personen- und untersuchungsbedingter Störvariablen.

4.2.3. Kontrollierte Störvariablen der Benutzer-System-Interaktion

Die große Zahl der mit dem Eintreten des Benutzers in die Testsituation zu berücksichtigenden Einflussfaktoren, stellt im Rahmen experimenteller Studien zum Informationssuchverhalten eine besondere Herausforderung dar, da sie die interne Validität der Ergebnisse beeinträchtigen kann. Um im Vorfeld einer Studie sicherzustellen, dass die interessierenden Handlungsaspekte auch angemessen beobachtet werden können, ist es daher notwendig, Maßnahmen zu ergreifen, um möglichst viele solcher Störgrößen auszuschließen. Wie in Abschnitt 4.1 bereits erwähnt, bietet das Laborexperiment den Vorteil, die Untersuchungsbedingungen systematisch verändern und kontrollieren zu können. Grundsätzlich unterscheidet man dabei zwei Arten von Störvariablen: *personenbezogene Störvariablen*, wie Vorwissen, Alter oder Geschlecht der Testpersonen und *un-*

tersuchungsbedingte Störvariablen, wie Lern- und Ermüdungseffekte. Der folgende Abschnitt stellt einige Methoden vor, die sich im IR-Kontext eignen, um den Einfluss von Störvariablen möglichst gering zu halten. Da die jeweilige Bedeutung der einzelnen personenbezogenen Einflussfaktoren für die Suchergebniswahrnehmung in Kapitel 2 bereits ausführlich dargestellt ist, beschränkt sich die folgende Darstellung auf eine Diskussion der verschiedenen Kontrolltechniken. Der Beitrag untersuchungsbedingter Einflussfaktoren hingegen steht in direktem Zusammenhang mit dem gewählten Untersuchungsdesign und wird deshalb an dieser Stelle ausführlicher erläutert. Abschnitt 4.2.3.3 diskutiert darüber hinaus Möglichkeiten zur Erfassung von Sucherfahrungen und Domänenwissen, da beide Einflussfaktoren im zweiten Experiment als Störvariablen geprüft werden. Die Berücksichtigung solcher erhobener Störfaktoren im Rahmen der statistischen Auswertung mit Hilfe einer Kovarianzanalyse wird hingegen in Abschnitt 4.3.2.3 erläutert.

4.2.3.1. Kontrolle personenbezogener Störvariablen

Laut Bortz und Döring (2006, S. 524) stellt die *Randomisierung* die wichtigste Technik zur Kontrolle personengebundener Störvariablen dar und kommt auch im Rahmen von IIR-Studien häufig zum Einsatz. In Bezug auf den Vergleich zweier unterschiedlicher Systeme hinsichtlich des Sucherfolgs, soll durch eine zufällige Zuordnung der Testteilnehmer bspw. verhindert werden, dass besonders viele Personen mit hohem oder niedrigem Vorwissen derselben Untersuchungsgruppe zugeteilt sind. Von Vorteil ist in diesem Zusammenhang, dass so bei ausreichender Stichprobengröße sowohl bekannte als auch unbekannte Störvariablen gleichmäßig auf die betrachteten Untersuchungsgruppen verteilt werden (ebd., S. 525). Die *Parallelisierung* ist eine weitere Kontrolltechnik, die bei personengebundenen Störvariablen angewendet werden kann. In diesem Fall erfolgt die Kontrolle der Störvariablen durch eine gezieltere Zuteilung der Teilnehmer zu den einzelnen Untersuchungsgruppen. Dazu wird bei der Zuordnung darauf geachtet, dass die Testpersonen in den unterschiedlichen Untersuchungsgruppen hinsichtlich einer oder mehrerer Störvariablen vergleichbar sind. Im IIR-Kontext kommt dieser Ansatz häufig in Bezug auf das Geschlecht der Testteilnehmer zum Einsatz. Eine Randomisierung innerhalb der Geschlechtergruppen sollte aber weiterhin vorgenommen werden, um bspw. das Vorwissen der Testteilnehmer weiterhin zu kontrollieren. Anderenfalls müsste dieses vor dem Experiment erfasst und die Probanden anschließend paarweise auf die betrachteten Untersuchungsgruppen verteilt werden. In diesem Fall würde jedoch ausschließlich das Vorwissen der Teilnehmer und keine anderen personengebunden Störvariablen kontrolliert. Eine Erhebung solcher Störvariablen kann jedoch trotzdem hilfreich sein, da sie es erlaubt, die entsprechenden Einflüsse mit Hilfe einer Kovarianzanalyse im Zuge der statistischen Auswertung aus den Daten herauszupartialisieren (vgl. Abschn. 4.3.2.3) Eine letzte Kontrolltechnik, auf die an dieser Stelle eingegangen werden soll, betrifft die Methode der Konstanthaltung. Dabei wird bei der Auswahl bzw. Zulassung der Testpersonen darauf geachtet, dass alle Versuchsteilnehmer sich in Bezug auf eine vermutete Störvariable nicht unterscheiden. Im Rahmen von IIR-Studien kann dies bspw. die Beschränkung auf eine bestimmte Altersgruppe oder die Auswahl ausschließlich weiblicher bzw. männlicher Testpersonen betreffen. Des Weiteren bietet sich die Methode der Konstanthaltung auch in Bezug auf untersuchungsbedingte Störvariablen an, die im nächsten Abschnitt diskutiert werden. In diesem Zusammenhang sollte bspw. darauf geachtet werden, die Testumgebung für alle Teilnehmer konstant zu halten, indem etwa Testinstruktionen verschriftlicht oder störende Umgebungsgeräusche

vermieden werden.

Zusammenfassend hebt die Diskussion der Stärken und Schwächen der einzelnen Kontrolltechniken die Wichtigkeit der Randomisierung im Rahmen experimenteller Studien zum Informationssuchverhalten hervor, um die interne Validität der Ergebnisse zu gewährleisten und sowohl bekannte als auch unbekannte Störvariablen kontrollieren zu können. Gerade dieser letzte Punkt ist angesichts der Tatsache, dass im Suchprozess eine Vielzahl zusätzlicher Einflussgrößen (kognitive Stile, Selbstwirksamkeitserwartungen, etc., vgl. Kap. 2) zu berücksichtigen sind, von großer Bedeutung.

4.2.3.2. Kontrolle untersuchungsbedingter Störvariablen

Wie für jede nutzerbasierte Studie, stellt sich auch bei interaktiven IR-Evaluierungen die Frage, ob tatsächlich natürliches Nutzerverhalten beobachtet werden kann. Aus der psychologischen Forschung sind eine Reihe von Mechanismen bekannt, die im Kontext von Testsituationen zu einer Verfälschung der Ergebnisse führen können. Einige dieser Effekte, die im IR-Kontext einem natürlichen Such- und Bewertungsverhalten entgegenstehen könnten, werden im Folgenden kurz erläutert. Im Wesentlichen lassen sich diese in zwei Gruppen einteilen, je nachdem ob es sich um Probleme im Suchprozess oder der anschließenden Datenerhebung (z.B. mit Hilfe von Frageitems) handelt.

Zur ersten Kategorie gehören Fehlerquellen wie *Lern-* und *Ermüdungseffekte*, die nach der Bearbeitung einer oder mehrerer Aufgaben auftreten und das Such- und Bewertungsverhalten der nächsten Aufgaben beeinflussen. Diese können z.B. dazu führen, dass im Laufe eines Experiments ein Leistungsanstieg durch Lernen bzw. ein Leistungsabfall durch Ermüdung eintreten kann. Clemmensen und Borlund (2016) beobachten bspw., dass die Testteilnehmer ihrer Studie bei der letzten Aufgabe signifikant aktiver werden, d.h. mehr Webseiten aufrufen und mehr Suchanfragen stellen. Ebenso können Ermüdungseffekte und Motivationsabsenkungen aber auch der Grund dafür sein, dass Testpersonen die Aufgaben schneller erledigen, mit dem Ziel, den Test möglichst früh zu beenden (Kelly, 2009, S. 52). Darüber hinaus können Wechselwirkungen zwischen Suchaufgaben und anderen Untersuchungsbedingungen auftreten (ebd., S. 52). Dies ist bspw. der Fall, wenn einzelne Aufgaben einen anderen Schwierigkeitsgrad aufweisen oder ein System für eine Domäne bessere Suchergebnisse zurückliefert (ebd., S. 52). Um derartige Topiceffekte auszuschließen, sollte deshalb die Reihenfolge der Aufgabenbearbeitung variiert werden. Sollen darüber hinaus auch unterschiedliche Systeme verglichen werden, ist weiterhin darauf zu achten, dass auch ihre Reihenfolge unabhängig von den Aufgaben variiert wird.

Eine Reihe von Studien beschäftigen sich dediziert mit Reihenfolgeeffekten im Rahmen von IR-Studien, die zur Verzerrung der Relevanzwahrnehmung der Teilnehmer führen können (Eisenberg u. Barry, 1988; Purgailis Parker u. Johnson, 1990; Huang u. Wang, 2004). Die Arbeit von Eisenberg und Barry (1988) umfasst zwei Experimente mit zwei verschiedenen Messverfahren, deren Messergebnisse anschließend verglichen werden. Während die Teilnehmer des einen Experiments die Relevanz der Dokumente auf einer kategorialen Skala mit sieben Abstufungen bewerten, erhalten die Teilnehmer des anderen Experiments die Anweisung, die Relevanz der Dokumente durch eine Zahl auszudrücken (Magnitude-Estimation). Die nach Relevanz sortierte Reihenfolge der aus 15 Dokumenten bestehenden Ergebnislisten (hoch-nach-niedrig vs. niedrig-nach-hoch) wird in beiden Experimenten als unabhängige Variable variiert. Insgesamt

zeigt sich, dass die Reihenfolge in der Tat das Antwortverhalten der Testpersonen beeinflusst. Sind die Dokumente in der hoch-nach-niedrig Reihenfolge sortiert, neigen die Teilnehmer dazu, die Relevanz der Dokumente zu unterschätzen. Im entgegengesetzten Fall ist eine systematische Überschätzung der Relevanz der Dokumente erkennbar (ebd., S. 295). Die Ergebnisse der Magnitude-Estimation fallen weniger eindeutig aus als die der kategorialen Skala, was bedeuten könnte, dass sich dieses Skalierungsverfahren weniger anfällig für derartige Effekte zeigt (ebd., S. 297 f.). Im Gegensatz dazu können Purgailis Parker und Johnson (1990) in einem ähnlichen Experiment keinen Reihenfolgeeffekt nachweisen, was die Autoren im Wesentlichen auf die zum Teil geringe Anzahl der präsentierten Dokumente (max. 15 Dokumente pro Ergebnisliste) zurückführen. Im Rahmen der von Purgailis Parker und Johnson (ebd.) durchgeführten Untersuchung hat jede Testperson die Aufgabe, eine eigene Suche auszuführen und die erhaltenen Treffer hinsichtlich ihrer Relevanz zu beurteilen, wobei für jede Testperson die Reihenfolge der Treffer randomisiert wird. Zur Operationalisierung der Relevanzurteile dient in dieser Studie eine 3-stufige kategoriale Skala mit den Ausprägungen *relevant*, *irrelevant* und *weiß nicht*. Darauf aufbauend gehen Huang und Wang (2004) der Frage nach, ob tatsächlich ein Zusammenhang zwischen der Anzahl der zu bewertenden Dokumente und dem Auftreten von Reihenfolgeeffekten bei der Relevanzbeurteilung besteht. Die Untersuchung umfasst zwei Phasen. In der ersten Phase bewerteten die Testteilnehmer 40 Dokumente in jeweils unterschiedlicher Reihenfolge. In der zweiten Phase, die in einem zeitlichen Abstand von zwei Monaten stattfindet, dienen die in Phase 1 getroffenen Relevanzurteile als Ausgangspunkt zur Sortierung der Ergebnislisten. Neben der nach Relevanz sortierten Reihenfolge (hoch-nach-niedrig vs. niedrig-nach-hoch) wird diesmal auch die Anzahl der zu bewertenden Dokumente als unabhängige Variable variiert. Die Messung der Relevanz erfolgt, wie in dem oben beschriebenen Experiment von Eisenberg und Barry (1988), in beiden Phasen auf einer kategorialen Bewertungsskala mit sieben Abstufungen. Insgesamt können die Ergebnisse von Eisenberg und Barry (ebd.) bestätigt werden. Darüber hinaus zeigt sich, dass dieser Effekt keine Rolle zu spielen scheint, wenn nur wenige (weniger als 15) oder sehr viele (mehr als 60) Dokumente bewertet werden müssen (Huang u. Wang, 2004, S. 974 ff.). Dies deckt sich mit den Beobachtungen von Purgailis Parker und Johnson (1990), die bei der Betrachtung von weniger als 15 Dokumenten ebenfalls keinen Reihenfolgeeffekt nachweisen können. Die Tatsache, dass bei einer hohen Anzahl von Dokumenten keine Reihenfolgeeffekte auftreten, lässt Huang und Wang (2004, S. 978) auf Ermüdungseffekte bei einem Teil der Probanden schließen.

Darüber hinaus können experimentelle Studien der unbewussten Einflussnahme durch den Testleiter unterliegen (Bortz u. Döring, 2006, S. 82). Dieser in der Sozialpsychologie *Versuchsleitereffekt* genannte Störeinfluss, bezeichnet die Beeinflussung von Untersuchungsergebnissen durch die soziale Interaktion zwischen Testleiter und Testperson. Neben der emotionalen Atmosphäre kann die Instruktion der Testpersonen die Art und Weise beeinflussen, in der diese die zu bearbeitenden Aufgaben erledigen. Um hier die Teilnehmer möglichst einheitlich zu informieren, kann die Einweisung z.B. schriftlich erfolgen, durch den Testleiter verlesen oder in Form eines Audiofiles bzw. Einführungsvideos geschehen. Um darüber hinaus auch im Laufe des Experiments zu vermeiden, dass die Erwartungen des Versuchsleiters die Reaktionsweisen der Testteilnehmer mitbestimmen, empfehlen Bortz und Döring (ebd.) Doppelblindstudien, bei denen neben

den Teilnehmern auch der Versuchsleiter keine Kenntnis über die Versuchsgruppenzugehörigkeit besitzt und somit auch keine unbewussten Hinweise geben kann.

In die Kategorie der Bewertungs- und Itemeffekte fallen unbewusste Verhaltensweisen auf Seiten der Testpersonen, die potenziell zu einer Verzerrung der Antworten führen. Ähnlich wie im Kontext der Suchaufgaben kann auch die Itemreihenfolge in einem Fragebogen die Antwort der Testteilnehmer beeinflussen. Die früher beantworteten Fragen führen zu einer Meinungsbildung, die Einfluss auf die folgenden Antworten ausübt. Aus diesem Grund kann es hilfreich sein, die Items nicht thematisch zu gliedern sondern im Gegenteil für eine ausreichende Durchmischung zu sorgen. Zwei weitere bekannte Effekte sind die sog. *Tendenz zur Mitte* und die *Akquieszenz*, also Zustimmungstendenz. Erstere bezeichnet das Bestreben von Testpersonen extreme Antwortmöglichkeiten zu vermeiden, bspw. aufgrund einer Unsicherheit mit dem eigenen Urteil (Jonkisz et al., 2008). Als Strategie diesem Effekt entgegenzuwirken wird daher der Verzicht auf eine mittlere, neutrale Antwortkategorie und die Einführung eines *weiß-nicht* Feldes empfohlen (ebd.). *Akquieszenz* hingegen beschreibt ein Verhalten, bei dem die Testpersonen dazu neigen, der präsentierten Itemaussage unabhängig ihres Inhalts zuzustimmen (Kelly et al., 2008a). Um solche Zustimmungstendenzen zu kontrollieren, können Items gleichen Inhalts, aber in negativer Formulierung in den Fragebogen mit aufgenommen werden um die Antworten a posteriori auf Konsistenz prüfen zu können (Jonkisz et al., 2008). Da Antworttendenzen besonders bei global formulierten Items zu beobachten sind, empfehlen Neugebauer und Porst (2001, S. 24) darüber hinaus die Konstruktion sehr spezifischer Frageitems.

Eine weitere Fehlerquelle liegt in der Natur der Fragebogenerhebung selbst. „Aus Sicht der Probanden wird das Ausfüllen von Tests oder Fragebögen als Kommunikation erlebt. Testpersonen wissen, dass sie anderen Menschen durch den Test etwas über sich mitteilen und machen sich Gedanken darüber, wer sie sind, was sie mitteilen wollen und was nicht, bei wem die Informationen ankommen, wie der Empfänger auf sie reagieren könnte und was mit ihnen geschieht.“(Bortz u. Döring, 2006, S. 232) Vor diesem Hintergrund erscheint es plausibel, dass Testpersonen versucht sein könnten ihre Außenwirkung aktiv zu beeinflussen. In der Tat ist in der Psychologie solch ein Verhalten unter dem Stichwort *sozial erwünschter Antworttendenzen* bekannt (Bortz u. Döring, 2006; Jonkisz et al., 2008). Dies umfasst sowohl Elemente der Selbst- als auch der Fremdtäuschung (Jonkisz et al., 2008). Im IR-Kontext könnte dieses Verhalten bspw. bei der Beschreibung der eigenen Sucherfahrung auftreten. Auch besteht gerade bei Suchmaschinenbewertungen die Gefahr, dass die Involviertheit der Probanden in den Suchprozess dazu führt, dass sie das Gefühl haben, nicht nur das System, sondern auch ihre eigene Leistung zu bewerten (vgl. Abschn. 3.3.1.3). In der Tat können Kelly et al. (2008b) nachweisen, dass die Teilnehmer ihr Zufriedenheitsurteil nach einer realistischen Rückmeldung zu ihrem erreichten Recall nach unten korrigieren (vgl. Abschn. 2.2.2). Des Weiteren ist in der genannten Studie zu beobachten, dass die Teilnehmer die Zufriedenheitsskala nicht vollständig ausnutzen, sondern in der großen Mehrheit eine überdurchschnittliche Zufriedenheit mit dem System erkennen lassen. Dies könnte als Hinweis darauf gedeutet werden, dass einige Teilnehmer sich dazu verpflichtet fühlen, bei der Abfrage der Zufriedenheit mit dem präsentierten Suchsystem eine möglichst hohe Bewertung auszudrücken, um den Testleiter nicht zu enttäuschen. Dieser Anforderungsdruck wird auch von Kelly et al. (2008a, S. 125) beschrieben: „In the context of interactive IR experiments, it seems

reasonable for subjects to interpret that desired effects are likely to translate into positive system ratings, which might suggest why subjects tend to inflate their ratings of systems. In addition, subjects may not want to offend the researcher by rating a system poorly.“ Jonkisz et al. (2008) empfehlen als Kontrollstrategie, die Anonymität der Befragung zu unterstreichen und deutlich zu machen, dass nicht die Leistung der Probanden, sondern die Leistung des Suchsystems im Vordergrund der Untersuchung steht. Weiterhin gibt es Hinweise, dass der Grad der von den Probanden wahrgenommenen Anonymität bei Onlinebefragungen im Vergleich zu Interviews und Papierfragebögen höher ausfällt (Richman et al., 1999, S. 756): „Also, if respondents type responses that seem to disappear into the computer, they may feel more anonymous than they do taking traditional tests, in which there is a concrete reminder of the evaluation, such as a printed questionnaire or interviewer.“ Allerdings könnte dieser Effekt im Zuge der öffentlichen Diskussion über Datenschutz und Privatsphäre in Zeiten von Big Data und Digitalisierung heute geringer ausfallen (Kelly et al., 2008a).

Zusammenfassend lässt sich sagen, dass der Übergang vom systemzentrierten zum nutzerzentrierten Paradigma zu einer erhöhten Komplexität und somit auch zu neuen Fehlerquellen führt. Dies hat auf der einen Seite den Nachteil, dass typische Probleme aus der experimentellen Psychologie, wie Reihenfolgeeffekte und Antworttendenzen berücksichtigt werden müssen. Auf der anderen Seite kann jedoch bei der Planung und Durchführung interaktiver IR-Experimente auf die Erfahrungen und Techniken aus diesem Bereich zurückgegriffen werden.

4.2.3.3. Erfassung von Sucherfahrungen und Domänenwissen

Nachdem in den vorangegangenen beiden Abschnitten verschiedene Möglichkeiten zur Kontrolle personen- und untersuchungsbedingter Störvariablen herausgearbeitet wurden, besteht das Ziel des folgenden Abschnitts darin zu untersuchen, in welcher Form und in welchem Umfang die Berücksichtigung einer konkreten Störgröße, nämlich des Vorwissens der Testpersonen, in experimentellen Kontexten erfolgen kann. Im Folgenden werden einige Operationalisierungsunterschiede der in Abschnitt 2.1.2 bereits inhaltlich vorgestellten Studien verglichen und im Hinblick auf ihre Eignung für den Einsatz in einem experimentellen Forschungsdesign bewertet.

Anhand der in Abschnitt 2.1.2 betrachteten Studien lassen sich drei verschiedene Herangehensweisen zur Erfassung des Vorwissens der Testteilnehmer identifizieren: das Ableiten der Erfahrung aus dem Verhalten der Testpersonen (Kelly u. Cool, 2002; White u. Morris, 2007), die Einstufung der Probanden basierend auf ihrer Selbstauskunft (Kissel, 1995; Hölscher u. Strube, 2000; Palmquist u. Kim, 2000; Richter et al., 2001b; Al-Maskari u. Sanderson, 2006) und die Konstanthaltung der Erfahrung (Vakkari u. Hakala, 2000; Wildemuth, 2004; Hölscher u. Strube, 2000; Jenkins et al., 2003; Dong et al., 2005). Logfileanalyse zeigen bspw., dass es bis zu einem gewissen Grad tatsächlich möglich ist, aus dem Verhalten der Nutzer Rückschlüsse auf ihre Sucherfahrung und ihr Domänenwissen zu ziehen (White u. Morris, 2007; Kelly u. Cool, 2002). Obwohl dieser Ansatz natürlich auch im Rahmen eines experimentellen Forschungsdesigns angewendet werden kann, da er keinen zusätzlichen Aufwand auf der Teilnehmerseite verursacht, bleibt doch eine Restunsicherheit, ob die Teilnehmer aufgrund ihres Suchverhaltens wirklich korrekt als Experten bzw. Anfänger identifiziert werden. Demgegenüber steht der Ansatz, das Domänenwissen und die Sucherfahrung der Teilnehmer explizit mit Hilfe von Fragebögen zu erfassen. Ein allgemeines und valides Instrument zur Erfassung von Computerwissen, Computerängst-

lichkeit und computerbezogenen Einstellungen stellt bspw. das Inventar zur Computerbildung (INCOBI) von Richter et al. (2001b) dar. Hier wird mit einer Mischung aus Wissens- und Quizfragen versucht, sowohl die theoretischen Kenntnisse der Probanden, als auch ihre praktische Vorerfahrung zu erfassen. Zusammen mit dem Fragebogen zum Internetwissen von Hölscher (2000) stellt INCOBI eine wichtige Inspirationsquelle für den in dieser Arbeit verwendeten Fragebogen zur Sucherfahrung der Probanden dar (vgl. Abschn. 6.3.3). Die größere Genauigkeit der selbstauskunftsbasierten Abfrage, wie sie in den meisten Studien verwendet wird, geht jedoch mit einem höheren Aufwand für die Testpersonen einher. Im Rahmen dieser Arbeit wird dieser zusätzliche Aufwand aber zu Gunsten einer gesicherteren Einordnung der Testteilnehmer in Kauf genommen. Der dritte Ansatz besteht in der Konstanthaltung der individuellen Vorerfahrung. In den Studien von Hölscher und Strube (2000), Jenkins et al. (2003) und Dong et al. (2005) geschieht dies durch das Hinzuziehen anerkannter Domänenexperten (z.B. Krankenschwestern mit einer speziellen Ausbildung für Patienten mit Osteoporose). Leider scheitert ein solcher Ansatz in experimentellen Settings häufig an der Verfügbarkeit entsprechender Experten, da hier in der Regel eine größere Anzahl an Probanden benötigt wird. In Längsschnittstudien liegt der Schwerpunkt darauf, Hypothesen über Einflussgrößen und Wirkungszusammenhänge mithilfe von Daten, die über einen längeren Zeitraum erhoben werden, empirisch zu prüfen. Dazu starten die Kursteilnehmer in den Studien von Vakkari und Hakala (2000) und Wildemuth (2004) zu Kursbeginn mit einem vergleichbaren Kenntnisstand und die Autoren untersuchen, in welchem Maße sich dieser im Laufe des Untersuchungszeitraums verändert. Diese Form der Konstanthaltung ist weniger schwierig umzusetzen und kann in Verbindung mit einer selbstauskunftsbasierten Abfrage vermutlich zu einer valideren Daten- und Interpretationsbasis beitragen. Aus diesem Grund werden im Rahmen der hier durchgeführten Untersuchungen bspw. aus IT-orientierten Studiengängen nach Möglichkeit nur Studienanfänger als Testpersonen rekrutiert, um so ihren Erfahrungsvorsprung gegenüber anderen Studierendengruppen gering zu halten.

Dieser Abschnitt hat gezeigt, dass im IR-Kontext im Wesentlichen drei Ansätze existieren, um das Vorwissen der Probanden zu berücksichtigen: die Verhaltensanalyse, die Selbstauskunft und die Konstanthaltung. Welche Methode im Einzelfall vorzuziehen ist, hängt jedoch von der jeweiligen Fragestellung und Schwerpunktsetzung des durchzuführenden Experiments ab. Im Rahmen dieser Arbeit finden sowohl Selbstauskunfts- und Fragebogentechniken als auch die Konstanthaltung der Teilnehmererfahrung Anwendung.

4.3. Angewendete statistische Verfahren

Der folgende Abschnitt stellt die den theoretischen Unterbau für die statistische Auswertung der im Rahmen dieses Forschungsvorhabens durchgeführten Experimente bereit. Da sowohl den zu untersuchenden Forschungshypothesen als auch dem gewählten Forschungsdesign eine Untersuchung des wechselseitigen Einflusses von Systemgüte und Erwartungshaltung auf die abhängigen Variablen zugrunde liegt, richtet sich das Hauptaugenmerk dabei auf einen Überblick über die verwendeten varianzanalytischen Methoden (vgl. Abschn. 4.3.2). Im Folgenden wird jedoch zunächst die im Kontext der Auswertung der Nutzerzufriedenheit verwendete Vorgehensweise zur Skalenbildung vorgestellt. Der letzte Abschnitt schließlich diskutiert die konkrete Umsetzung der Auswertung mit Hilfe der Statistiksoftware R.

4.3.1. Skalenbildung für Zufriedenheitsitems

Viele wissenschaftliche Fragestellungen in der Psychologie oder den Sozialwissenschaften beziehen sich auf Konzepte, die einer direkten Messung nicht zugänglich sind. Wie in Abschnitt 4.2.2.3 dargestellt, trifft dies auch auf die Benutzerzufriedenheit zu, deren Erfassung typischerweise über Fragebögen realisiert wird. Darüber hinaus enthält Zufriedenheit, als komplexes Konstrukt, unterschiedliche Dimensionen, die mit jeweils eigenen Frageitems adressiert werden müssen. Dabei stehen im Kontext von IR-Studien jedoch nicht die Antworten der Testpersonen auf spezielle Fragebogenitems im Vordergrund, sondern die Zufriedenheit der Testpersonen mit allgemeineren Konzepten, wie Benutzerfreundlichkeit, inhaltliche Relevanz und Systemgüte. Faktoranalytische Methoden erlauben es, Fragebogenitems nach ihrem Korrelationsgrad zu gruppieren und auf diese Weise solche Dimensionen zu identifizieren. Allerdings ist zu beachten, dass hier einzig die relative Korrelationsstärke zwischen den Items berücksichtigt wird, die inhaltliche Interpretation dieser Gruppierung hingegen ist durch den Forscher vorzunehmen. In diesem Sinne können Faktoranalysen als Verfahren interpretiert werden, bei denen aufgrund von Messdaten Forschungshypothesen formuliert werden, welche dann wiederum in Experimenten zu überprüfen sind. Darüber hinaus erlauben sie es jedoch auch, die Information vieler miteinander korrelierender Frageitems in einige wenige Faktoren zusammenzufassen, die dann wiederum mit Hilfe varianzanalytischer Methoden auf ihre Abhängigkeit von den unabhängigen Variablen überprüft werden können. Es findet also eine Kondensation in wenige Kennzahlen statt, die im Idealfall zu einer stärkere Variation zwischen den einzelnen Untersuchungsgruppen führt. Im Rahmen dieser Arbeit liegt das Hauptaugenmerk auf diesem skalenbildenden explorativen Aspekt der faktoranalytischen Methoden. Grundsätzlich ist der Term Faktoranalyse als Oberbegriff für eine ganze Reihe verschiedener Verfahren zur Bündelung von Fragebogenitems zu sehen (Bortz, 2005, S. 514 ff.) die folgende Darstellung beschränkt sich jedoch auf die im Rahmen dieser Arbeit gewählte Hauptkomponentenanalyse (Principal-Component-Analysis (PCA)).

Wie bereits angedeutet besteht das Ziel einer PCA darin, aus den gegebenen Frageitems eine reduzierte Anzahl von Skalen zu extrahieren, die jeweils unabhängige Aspekte der Fragebogendaten beschreiben, zusammengenommen jedoch immer noch die vollständige Information über die in den Antwortdaten enthaltenen Unterschiede repräsentieren. Der Begriff Skala ist in diesem Zusammenhang als gewichtete Summe von Itemwerten zu verstehen und wird häufig auch synonym als Faktor bezeichnet. Im Kontext dieser Arbeit erwartet man bspw., dass Items zusammengefasst werden können, die unterschiedliche Dimensionen der Nutzerzufriedenheit, wie Gebrauchstauglichkeit oder Suchleistung abfragen. Das Ziel ist es also, Gruppen stark miteinander korrelierender Frageitems zu identifizieren und diese jeweils in einem gemeinsamen Faktor zu bündeln. Gleichzeitig sollte jedes Frageitem nur in wenigen Faktoren enthalten sein. Dies stellt sicher, dass die Korrelationen zwischen den abgeleiteten Skalen möglichst gering ausfallen und im Idealfall jeder Faktor eine unabhängige Dimension der Zufriedenheit erfasst. Ähnlich einer Varianzanalyse sollen die gewählten Skalen darüber hinaus einen Großteil der in den Fragebogendaten vorhandenen Varianz aufklären, um die in den Daten beobachteten Unterschiede möglichst unverfälscht wiederzugeben. Eine Hauptkomponentenanalyse erlaubt es nun, Faktoren mit genau diesen Eigenschaften zu berechnen, indem aus der Menge der stark miteinander korrelierenden Items sukzessive die Gruppe mit der höchsten Varianz zu einem Faktor zusammen-

gefasst wird. Mathematisch gesehen entspricht dies einer geeigneten Koordinatentransformation der zu den Fragebogendaten korrespondieren Kovarianzmatrix (Bortz, 2005, S. 518 ff.). Vor der Durchführung einer PCA sind einige Voraussetzungen zu überprüfen, um die Validität der extrahierten Faktoren sicherzustellen. Zunächst sollten die erhobenen Daten intervallskaliert sein, damit die Addition verschiedener Itemwerte sinnvoll ist. Weiterhin muss sichergestellt werden, dass eine im Vergleich zur Itemanzahl ausreichende Stichprobengröße vorliegt. Dies kann mit Hilfe des Kaiser-Meyer-Olkin (KMO)-Kriteriums überprüft werden (Field et al., 2012, S. 776 f.). Dabei gelten Werte zwischen $0{,}5$ und $0{,}7$ als akzeptabel, Werte zwischen $0{,}7$ und $0{,}8$ als gut, Werte zwischen $0{,}8$ und $0{,}9$ als sehr gut und Werte größer als $0{,}9$ als hervorragend. Darüber hinaus müssen die Frageitems einerseits ausreichend miteinander korrelieren, andererseits können auch Multikollinearitäten, d.h. zu starke Korrelationen zwischen einzelnen Items zu Problemen führen. Ersteres kann durch einen Bartlett-Test getestet werden, der signifikante Korrelationen anzeigt. Multikollinearität kann durch die Determinante der Korrelationsmatrix überprüft werden, die einen Wert von $0{,}00001$ nicht unterschreiten sollte (ebd., S. 771).

Prinzipiell würde eine Hauptkomponentenanalyse ebenso viele Faktoren generieren, wie Frageitems in die Analyse eingehen, die jedoch eine immer geringere Varianzaufklärung aufweisen. Im Sinne einer Reduktion der Daten auf wenige aussagekräftige Faktoren stellt sich somit die Frage, wie viele Faktoren sinnvollerweise extrahiert werden sollten. Dazu gibt es in der Literatur eine Reihe empirischer Verfahren, die entweder auf einer graphischen Analyse (Scree-Test) oder der Auswertung spezieller Kennzahlen (Eigenwertkriterium, Parallelanalyse, VSS-Kriterium, MAP-Kriterium) beruhen (Luhmann, 2013, S. 290 ff.). Da alle diese Kriterien Raum für Interpretation lassen, wird häufig, und auch im Rahmen dieser Arbeit, auf eine Kombination dieser Techniken zurückgegriffen. An dieser Stelle wird noch einmal deutlich, dass explorative Faktoranalysen als Werkzeug zur Hypothesenbildung anzusehen sind, da die Ergebnisse, wie hier die gewählte Faktorenzahl, von der Interpretation des Ausführenden abhängig sind.

Nach der Extraktion der gewählten Anzahl von Faktoren mit Hilfe einer PCA ist eine weitere Koordinatentransformation notwendig, um zu einer interpretierbaren Faktorstruktur zu gelangen. Typischerweise sind die Frageitems in mehreren der mit Hilfe der PCA bestimmten Faktoren enthalten, was die inhaltliche Interpretation der Faktoren erschweren kann. Man spricht in diesem Zusammenhang auch von Doppel- bzw. Mehrfachladung eines Frageitems. Eine weitere Rotation der Faktoren kann hier Abhilfe schaffen, wobei zwischen orthogonalen und obliquen Rotationen unterschieden wird, je nachdem ob eine Korrelation der Faktoren zu erwarten ist (Bortz, 2005, S. 547 ff.). Erhält man an dieser Stelle keine zufriedenstellende Lösung, kann eine Reduktion oder Erweiterung der Faktoren und Frageitems getestet werden, um eine interpretierbare Faktorlösung zu erhalten. Nach Abschluss dieses Rotationsschrittes wird überprüft, welche Frageitems vorrangig auf einen einzelnen Faktor laden und diese pro Faktor zu einer Skala kombiniert (Field et al., 2012, S. 793 ff.). Im letzten Schritt werden Reliabilität und Validität der finalen Faktoren analysiert. Dabei wird die Reliabilität anhand der internen Konsistenz mit Cronbachs Alpha überprüft. Zur Überprüfung der Validität hingegen wird ein Außenkriterium verwendet, dass im Rahmen dieser Arbeit aus zwei zusätzlichen Frageitems nach der allgemeinen Zufriedenheit der Testteilnehmer besteht.

Zusammenfassend stellt dieser Abschnitt somit einen Leitfaden zur Skalenbildung für die im

experimentellen Teil dieser Arbeit verwendeten Zufriedenheitsfragebögen bereit. Eine Übersicht über die Umsetzung mit der Statistiksoftware R ist darüber hinaus in Abschnitt 4.3.3 zu finden. Die konkrete Umsetzung im Rahmen der Nutzerstudien kann hingegen den Abschnitten 6.4.4 und 7.4.4 entnommen werden.

4.3.2. Varianzanalytische Auswertungsmethodik

Wie in Abschnitt 4.1 dargestellt, erlauben es Laborstudien, den Einfluss einzelner unter der Kontrolle des Experimentators stehender unabhängiger Variablen auf die im Experiment gemessenen Parameter zu untersuchen. Im Rahmen dieser Arbeit handelt es sich dabei typischerweise um den Einfluss von Systemleistung und Benutzererwartung auf Suchleistung und Zufriedenheit der Teilnehmer. Zur Überprüfung der untersuchten Forschungshypothesen werden jedoch Verfahren benötigt, die es ermöglichen, die Unterschiede zwischen den einzelnen Untersuchungsgruppen statistisch valide zu quantifizieren. Varianzanalytische Methoden ermöglichen es, Gruppenunterschiede auch beim Vorhandensein mehrerer unabhängiger Variablen zu untersuchen und können in diesem Sinne als Erweiterung der klassischen T-Tests gesehen werden, die den Vergleich zweier Untersuchungsgruppen ermöglichen. Damit stellen sie ein geeignetes Mittel für die Auswertung der in Abschnitt 4.1.2 vorgestellten Forschungsdesigns dar. Die folgenden Abschnitte widmen sich zunächst unterschiedlichen Aspekten der varianzanalytischen Auswertung, bevor im letzten Abschnitt noch einmal auf die konkrete Umsetzung mit Hilfe der Statistiksoftware R eingegangen wird. Die folgende Diskussion orientiert sich, auch in Bezug auf die Notation, an den Darstellungen in (Bortz, 2005; Field et al., 2012).

4.3.2.1. Einfaktorielle Varianzanalyse

Im Folgenden wird ein kurzer Überblick über die Theorie und Praxis von Varianzanalysen bzw. Analysis of Variances (ANOVA) ohne Messwiederholung gegeben. Diese Verfahren eignen sich gerade für die Auswertung von Between-Subjects-Versuchsplänen, wie sie in Abschnitt 4.1.2.2 vorgestellt werden. Der Schwerpunkt liegt dabei auf den zugrundeliegenden Prinzipien anstelle einer genauen mathematischen Herleitung für die an dieser Stelle auf entsprechende Lehrbuchliteratur verwiesen wird (Bortz, 2005; Field et al., 2012).

Um die Herangehensweise an die Daten zu veranschaulichen, wird zunächst der Fall einer einfaktoriellen Varianzanalyse behandelt, bei der nur eine einzelne unabhängige Variable vorliegt. Um die Darstellung konkreter zu machen, wird im Folgenden speziell der Fall einer experimentellen Studie zum Informationssuchverhalten betrachtet, bei der die Systemgüte über drei unterschiedliche Leistungsabstufungen (gut (S_G), mittel (S_M), schlecht (S_S)) variiert wird. Die untersuchte Forschungshypothese wäre somit, dass eine bessere Systemleistung zu einer höheren Suchleistung der Testteilnehmer führt. Weiterhin wird angenommen, dass jeweils $n = 20$ Testpersonen mit dem jeweiligen Suchsystem arbeiten und anschließend die Systemleistung anhand der Benutzerprecision (BP) ermittelt wird. Somit bestehen die Untersuchungsdaten aus den Messwerten der insgesamt $3 \times 20 = 60$ Teilnehmer. Ziel ist es nun, die Forschungshypothese (H_1), dass die Suchleistung innerhalb der drei Treatmentgruppen mit steigender Systemleistung zunimmt, mit der Nullhypothese (H_0), dass die Systemleistung keinen Einfluss auf die Suchleistung hat, zu vergleichen. Die prinzipielle Idee einer Varianzanalyse ist es also, zu testen, ob sich die mittleren Suchleistungen der drei Untersuchungsgruppen signifikant voneinander unterscheiden (Bortz,

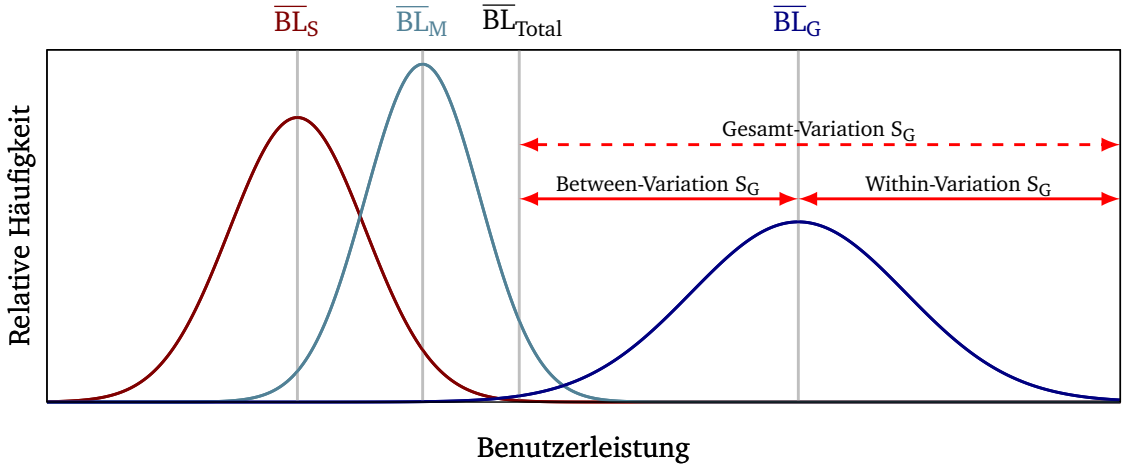


Abb. 4.6.: Beispielhafte Darstellung der Within- und Between-Gruppenvariation eines einfaktoriellen Between-Subjects-Designs. In einem IR-Experiment wird die Systemgüte über drei unterschiedliche Leistungsabstufungen (gutes System (S_G), mittleres System (S_M) u. schlechtes System (S_S)) variiert. Gezeigt ist die Verteilung der Benutzerleistung innerhalb der drei Versuchsgruppen. $\overline{BL}_S$, $\overline{BL}_M$, $\overline{BL}_G$ kennzeichnen die jeweiligen Gruppenmittelwerte der Benutzerleistung, während $\overline{BL}_{Total}$ den Mittelwert der Gesamtstichprobe bezeichnet. Weiterhin sind für die Untersuchungsgruppe S_G Gesamtvariation, Within-Variation und Between-Variation schematisch gekennzeichnet.

2005). Eine fiktive Verteilung der Benutzerleistung aufgeschlüsselt nach Treatmentgruppen ist in Abbildung 4.6 dargestellt. Aus der Graphik lässt sich qualitativ ablesen, dass eine bessere Systemgüte tatsächlich zu einer besseren Benutzerleistung zu führen scheint. Um diese Unterschiede quantifizieren zu können, wird überprüft welcher Anteil der beobachteten Varianz in der Stichprobe dem Einfluss der unabhängigen Variable zugerechnet werden kann und welcher Anteil durch die individuellen Voraussetzungen und andere Einflüsse erklärt werden muss. Beide Varianzanteile sind schematisch für die Untersuchungsgruppe S_G in Abbildung 4.6 gekennzeichnet, die notwendigen Rechenschritte sind hingegen in Tabelle 4.5 zusammengefasst. Daraus geht hervor, dass die Gesamtvariation der Messwerte, d.h. die Summe der quadratischen Abweichungen (QA_{Total}) vom Gesamtmittelwert der Stichprobe über alle drei Untersuchungsgruppen hinweg, aufgeteilt wird in einen Anteil der dem Treatment zuzuordnen ist ($QA_{Treatment}$), sowie einen Fehlerterm (QA_{Fehler}), der die Streuung der Daten durch nicht im Experiment kontrollierte Einflüsse, wie bspw. das Vorwissen der einzelnen Teilnehmer und ihre individuelle Suchgeschwindigkeit, beschreibt. Mathematisch stellt sich dieser Zusammenhang als

$$QA_{Total} = QA_{Treatment} + QA_{Fehler} \tag{4.2}$$

dar (Field et al., 2012, S. 410). Ist die beobachtete Variation zwischen den Untersuchungsgruppen ($QA_{Treatment}$) groß im Vergleich zu der Variation innerhalb der Untersuchungsgruppen (QA_{Fehler}), kann die Nullhypothese, dass sich die Gruppenmittelwerte nicht unterscheiden, verworfen werden. In diesem Fall erklärt die unabhängige Variable also zu einem guten Teil die gemessenen Unterschiede zwischen den Untersuchungsgruppen und man spricht von einer hohen Varianzaufklärung durch die unabhängige Variable. Um nun diese Aussage statistisch zu überprüfen,

Tab. 4.5.: Berechnungsschritte einer einfaktoriellen Varianzanalyse. p bezeichnet die Anzahl der Untersuchungsgruppen, n die Anzahl der Versuchspersonen pro Gruppe. In dem im Text genannten Beispiel zur Systemgüte gilt $p = 3$ (S_G, S_M, S_S) und $n = 20$, was auf $n \cdot p = 60$ Datensätze führt. Die Benutzerleistung von Testperson i in der Treatmentgruppe a wird als $BL_{i,a}$ bezeichnet.

Mittelwerte		**Quadratische Abweichungen (QA)**	
Gesamt	$\overline{BL} = \frac{1}{n \cdot p} \sum_i \sum_a BL_{i,a}$	Gesamt	$QA_{Total} = \sum_i \sum_a \left(BL_{i,a} - \overline{BL}\right)^2$
Gruppe	$\overline{BL}_a = \frac{1}{n} \sum_i BL_{i,a}$	Gruppen–Gesamt	$QA_{Treatment} = n \cdot \sum_a \left(\overline{BL}_a - \overline{BL}\right)^2$
		Innerhalb Gruppen	$QA_{Fehler} = \sum_i \sum_a \left(BL_{i,a} - \overline{BL}_a\right)^2$
Varianzen		**F-Test**	
Treatment	$\sigma^2_{Treatment} = \frac{QA_{Treatment}}{p-1}$	Treatment	$F_{Treatment} = \frac{\sigma^2_{Treatment}}{\sigma^2_{Fehler}}$
Fehler	$\sigma^2_{Fehler} = \frac{QA_{Fehler}}{p \cdot (n-1)}$		

werden die entsprechenden Varianzen $\sigma^2_{Treatment}$ und σ^2_{Fehler} mit Hilfe eines F-Tests verglichen (vgl. Tab. 4.5). Ist dieser Test signifikant, d.h. ist die Varianz zwischen den Untersuchungsgruppen ausreichend groß im Vergleich zur Varianz innerhalb der Untersuchungsgruppen, kann die Nullhypothese, dass alle Gruppenmittelwerte identisch sind, verworfen werden. Es ist allerdings zu beachten, dass eine signifikante Varianzanalyse nur bedeutet, dass signifikante Mittelwertunterschiede vorhanden sind, jedoch keine Information darüber liefert, welche der Mittelwerte sich unterscheiden. In dem betrachteten Beispiel muss also über einen weiteren sog. Posthoc-Test geklärt werden, welche der drei Systemgüten S_G, S_M und S_S sich signifikant unterscheiden. Des Weiteren können durch die Verwendung orthogonaler Kontraste in der Varianzanalyse gewichtete Mittelwertvergleiche überprüft werden (Bortz, 2005, S. 266 ff.).

Zwar wird an dieser Stelle auf eine ausführliche mathematische Herleitung der einfaktoriellen Varianzanalyse verzichtet, es ist jedoch notwendig, auf die impliziten Annahmen einzugehen, die dabei gemacht werden, um die Rahmenbedingungen zu verstehen, unter denen die Methode valide angewendet werden kann (ebd.). Die Zerlegung der Gesamtvariation in einen Treatment und Fehlerterm ist ohne weitere Annahmen gültig, für die Anwendung des F-Tests auf das Verhältnis der zugehörigen Varianzen müssen jedoch im Wesentlichen die folgenden Bedingungen überprüft werden (ebd., S. 284 ff.):

Skalenniveau – Die abhängige Variable muss mindestens intervallskaliert sein, damit Gruppenmittelwerte gebildet werden können (Field et al., 2012).

Normalverteilung – Die Residuen, d.h. die Abweichungen der Testwerte vom jeweiligen Gruppenmittelwert, müssen in jeder der betrachteten Treatmentgruppen einer Normalverteilung folgen. Die Überprüfung kann durch einen Test auf Normalverteilung, wie bspw. den Shapiro-Wilks-Test, erfolgen.

Varianzhomogenität – Die Fehlervarianz der Gesamtpopulation wird als Mittelwert der Varianzen der einzelnen Untersuchungsgruppen geschätzt. Aus diesem Grund dürfen sich die Varianzen der einzelnen Treatmentgruppen nicht signifikant voneinander unterscheiden, müssen

also über alle Untersuchungsbedingungen homogen sein. Diese Hypothese lässt sich bspw. über einen Bartlett- oder Levene-Test überprüfen (Bortz, 2005, S. 285 f.).

Unabhängigkeit der Fehlerkomponenten – Die Fehlerkomponente eines gegebenen Messwerts darf nicht von den Fehlerkomponenten eines anderen Messwerts abhängen. Diese Bedingung kann als erfüllt angesehen werden, wenn die Teilnehmer den Untersuchungsgruppen zufällig zugeordnet werden und jeder Teilnehmer nur einem Treatment ausgesetzt ist. Für den Fall, dass mehrere Messergebnisse pro Testperson miteinander vergleichen werden sollen, müssen hingegen Messwiederholungs- oder gemischte Designs verwendet werden, auf die in Abschnitt 4.3.2.4 eingegangen wird.

Bevor eine einfaktorielle Varianzanalyse verlässlich angewendet werden kann, müssen also die zugrundeliegenden Daten in Bezug auf die hier dargestellten Voraussetzungen hin überprüft werden. Im Allgemeinen gelten ANOVAs bei ausreichend großer Stichprobengröße ($n > 10$) und ausgeglichenen Versuchsplänen, d.h. der gleichen Anzahl von Testpersonen pro Untersuchungsgruppe, als robust gegenüber Verletzungen der Normalverteilungs- und Varianzhomogenitätsvoraussetzung (ebd., S. 287). Vorsicht ist allerdings bei kleinen Stichprobengrößen sowie großen Unterschieden in den Stichprobenumfängen zwischen den einzelnen Treatmentgruppen geboten. In diesen Fällen sollte auf nicht-parametrische, d.h. voraussetzungsfreie Tests zurückgegriffen werden, die in Abschnitt 4.3.3 behandelt werden.

Nach diesem Überblick über die Theorie und Praxis einfaktorieller ANOVAs, geht der folgende Abschnitt auf die für das hier angestrebte Forschungsvorhaben relevante Methodik der zweifaktoriellen Varianzanalyse ein.

4.3.2.2. Zweifaktorielle Varianzanalyse

Die Methodik der einfaktoriellen Varianzanalyse lässt sich auf den Fall mehrerer unabhängiger Variablen verallgemeinern, was auch als multivariate ANOVA bezeichnet wird. Sie ist damit auch für die Auswertung von den in Abschnitt 4.1.2.2 vorgestellten Between-Subjects-Designs geeignet, bei denen mehr als eine unabhängige Variable variiert wird. Da mit Systemleistung und Benutzererwartung im Rahmen dieser Arbeit zwei unabhängige Variablen vorliegen, beschränkt sich die Darstellung auf den zweifaktoriellen Fall.

Im Folgenden sei weiterhin konkret der Fall der beiden unabhängigen Variablen Systemgüte (S) und Benutzererwartung (E) mit jeweils zwei Ausprägungen (gutes System S_G u. schlechtes System S_S bzw. hohe Erwartung E_H u. niedrige Erwartung E_N) betrachtet. Dies führt auf einen Versuchsplan mit den $2 \times 2 = 4$ unterschiedlichen Untersuchungsgruppen (S_G/E_H, S_G/E_N, S_S/E_H, S_S/E_N). Das Vorgehen entspricht nun im Wesentlichen dem Fall der einfaktoriellen Varianzanalyse, bei der die Gesamtvariation (QA_{Total}) in den Testdaten zunächst additiv in einen Treatment ($QA_{Treatment}$) sowie einen Fehleranteil (QA_{Fehler}) zerlegt wird. Wie in Tabelle 4.6 dargestellt, werden nun jedoch in einem weiteren Schritt die Treatmentquadratabweichungen weiter aufgespalten. Dabei wird danach unterschieden ob die Variation durch die Systemgüte (QA_S), die Erwartungshaltung (QA_E) oder eine Wechselwirkung zwischen den beiden unabhängigen Variablen ($QA_{S\times E}$) aufgeklärt wird. Die genauen mathematischen Definitionen der einzelnen Anteile können Tabelle 4.6 entnommen werden. Relevant ist, dass analog zu Gleichung (4.2) der

Tab. 4.6.: Berechnungsschritte für eine zweifaktorielle Varianzanalyse mit den Faktoren Systemgüte (S) und Erwartungshaltung (E). p_S und p_E bezeichnen die Anzahl der jeweiligen Faktorstufen, was auf eine Untersuchungsgruppenanzahl von $p_S \cdot p_E$ führt. n gibt die Anzahl der Versuchspersonen pro Gruppe an. In dem im Text genannten Beispiel zur Benutzerzufriedenheit gilt $p_S = p_E = 2$. Die Benutzerzufriedenheit von Testperson i in der Treatmentgruppe mit Systemgüte a und Erwartungshaltung b wird als $Z_{i,a,b}$ bezeichnet.

Mittelwerte		**Quadratische Abweichungen (QA)**	
Gesamt	$\overline{Z} = \frac{1}{n \cdot p_S \cdot p_E} \sum_i \sum_a \sum_b Z_{i,a,b}$	Gesamt	$QA_{Total} = \sum_i \sum_a \sum_b \left(Z_{i,a,b} - \overline{Z}\right)^2$
System	$\overline{Z}_a = \frac{1}{p_E \cdot n} \sum_i \sum_b Z_{i,a,b}$	System–Gesamt	$QA_S = n \cdot p_E \sum_a \left(\overline{Z}_a - \overline{Z}\right)^2$
Erwartung	$\overline{Z}_b = \frac{1}{p_S \cdot n} \sum_i \sum_a Z_{i,a,b}$	Erwartung–Gesamt	$QA_E = n \cdot p_S \cdot \sum_b \left(\overline{Z}_a - \overline{Z}\right)^2$
System–Erwartung	$\overline{Z}_{a,b} = \frac{1}{n} \sum_i Z_{i,a,b}$	Interaktion	$QA_{S\times E} = n \cdot \sum_a \sum_b \left(\widetilde{Z}_{a,b} - \overline{Z}_{a,b}\right)^2$
Interaktion	$\widetilde{Z}_{a,b} = \overline{Z}_a + \overline{Z}_b - \overline{Z}$	Innerhalb Gruppen	$QA_{Fehler} = \sum_i \sum_a \sum_b \left(Z_{i,a,b} - \overline{Z}_{a,b}\right)^2$
Varianzen		**F-Test**	
System	$\sigma_S^2 = \frac{QA_S}{p_S - 1}$	System	$F_S = \frac{\sigma_S^2}{\sigma_{Fehler}^2}$
Erwartung	$\sigma_E^2 = \frac{QA_E}{p_E - 1}$	Erwartung	$F_E = \frac{\sigma_E^2}{\sigma_{Fehler}^2}$
Interaktion	$\sigma_{S\times E}^2 = \frac{QA_{S\times E}}{(p_S - 1)(p_E - 1)}$	Interaktion	$F_{S\times E} = \frac{\sigma_{S\times E}^2}{\sigma_{Fehler}^2}$
Fehler	$\sigma_{Fehler}^2 = \frac{QA_{Fehler}}{p_{S\cdot E} \cdot (n-1)}$		

Zusammenhang

$$\begin{aligned} QA_{Total} &= QA_{Treatment} + QA_{Fehler} \\ &= QA_S + QA_E + QA_{S\times E} + QA_{Fehler} \end{aligned} \tag{4.3}$$

besteht. Entscheidend ist, dass neben den sog. Haupteffekten QA_S und QA_E auch eine mögliche Interaktion der beiden unabhängigen Variablen mit berücksichtigt werden kann. Wie aus Tabelle 4.6 hervorgeht, enthält $QA_{S\times E}$ die Anteile der Treatmentvariation, die nicht durch Systemgüte und Erwartungshaltung allein erklärt werden können. Dies tritt dann ein, wenn die Wirkung des einen Faktors von der Faktorstufe der zweiten unabhängigen Variable abhängt. Das in Abschnitt 3.3.1.1 beschriebene C/D-Paradigma postuliert gerade solche einen Zusammenhang: Die Übererfüllung bzw. Enttäuschung von Erwartungen sollte zu einer größeren bzw. geringeren Zufriedenheitsreaktion führen. Hier hängt die Reaktion auf die Systemgüte also gerade von der Erwartungshaltung der Testpersonen ab. Um eine Aussage darüber treffen zu können, ob einer der beiden Haupteffekte einen Einfluss auf die abhängige Variable hat oder ob ggf. eine signifikante Wechselwirkung besteht, werden analog zur einfaktoriellen Varianzanalyse F-Tests durchgeführt. Diese beruhen wie im einfaktoriellen Fall auf dem Verhältnis zwischen den in Tabelle 4.6 definierten Varianzen σ_S^2, σ_E^2 und $\sigma_{S\times E}^2$ und der Varianz innerhalb der Gruppen σ_{Fehler}^2. Insgesamt sind also drei F-Tests, je einer für die Systemgüte, die Erwartungshaltung sowie deren Interaktion, durchzuführen (ebd., S. 297 f.).

Liegt keine signifikante Wechselwirkung vor, können ggf. signifikante Haupteffekte in Be-

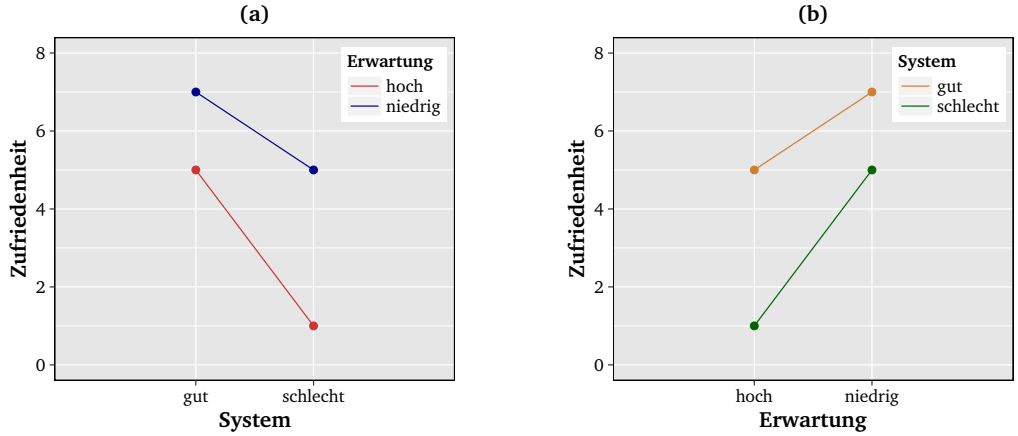


Abb. 4.7.: Beispielhafte Darstellung eines Wechselwirkungsdiagramms für ein zweifaktorielles Untersuchungsdesign. Dargestellt ist eine mit dem C/D-Paradigma im Einklang stehende Interaktion zwischen Systemgüte und Erwartungshaltung. Bild (a) zeigt die Zufriedenheit in Abhängigkeit der Systemgüte getrennt nach Erwartungshaltung, während Bild (b) die Zufriedenheit in Abhängigkeit der Erwartungshaltung getrennt nach Systemgüte zeigt. Das Enttäuschen bzw. Übererfüllen von Erwartungen führt jeweils zu niedriger bzw. höherer Zufriedenheit. Des Weiteren scheint der Unterschied zwischen gutem und schlechtem System bei hoher Erwartungshaltung größer auszufallen.

zug auf Systemgüte und Erwartungshaltung unabhängig und analog zur einfaktoriellen ANOVA interpretiert werden. Bei lediglich zwei Merkmalsausprägungen, wie in dieser Arbeit der Fall, entfällt auch die Notwendigkeit weiterer Posthoc-Tests, da jeweils nur zwei Gruppenmittelwerte miteinander zu vergleichen sind. Wird hingegen die Interaktion zwischen Systemleistung und Erwartungshaltung signifikant, können die Einflüsse der Haupteffekte nicht mehr unabhängig voneinander betrachtet werden (Bortz, 2005, S. 299). Um die Interpretation der Wechselwirkung zu erleichtern, wird auf sog. *Interaktionsdiagramme* zurückgegriffen, welche die Gruppenmittelwerte der abhängigen Variable einmal in Abhängigkeit der Systemleistung getrennt nach Erwartungshaltung und einmal in Abhängigkeit der Erwartungshaltung getrennt nach Systemleistung darstellen. Abbildung 4.7 zeigt solch ein Interaktionsdiagramm für die Ergebnisse einer fiktiven Nutzerstudie, für einen dem C/D-Paradigma entsprechenden Zusammenhang zwischen Nutzerzufriedenheit, Systemgüte und Erwartungshaltung. Ein Indiz für das Vorliegen einer Interaktion ist die Tatsache, dass die Linien in beiden Diagrammen nicht parallel verlaufen, der Einfluss sich also abhängig vom Wert der anderen unabhängigen Variable ändert. Bei der speziellen Form dieses Diagramms bei dem beide Linien jeweils dem gleichen Trend folgen (höhere Systemleistung entspricht jeweils einer höheren Zufriedenheit) spricht man auch von einer ordinalen Interaktion (ebd., S. 301). Zeigen die Linien hingegen in einem der Diagramme ein gegenläufiges Verhalten liegt eine hybride Interaktion vor (ebd., S. 301). Ist ein beiden Diagrammen ein gegenläufiges Verhalten zu beobachten, handelt es sich um eine disordinale Interaktion (ebd., S. 301).

Wie im Fall der einfaktoriellen ANOVA mit mehr als zwei Faktorstufen müssen ggf. Posthoc-Tests verwendet werden, um bei signifikanter Interaktion die Untersuchungsgruppen zu bestimmen, deren Mittelwerte sich in signifikanter Weise unterscheiden. Die im vorherigen Abschnitt für die einfaktorielle ANOVA genannten statistischen Voraussetzungen müssen auch im Fall der zweifaktoriellen Varianzanalyse erfüllt sein, um valide Ergebnisse zu erhalten (ebd., S. 328).

Dabei sind die Varianzhomogenität und die Normalverteilungsvoraussetzung nun jedoch über aller Untersuchungsgruppen, d.h. Kombinationen der unabhängigen Variablen, sicherzustellen. Jedoch gelten auch die Aussagen zur Robustheit des F-Tests bei ausreichend großer Stichprobengröße und balancierten Stichproben. Robuste Verfahren zur Durchführung einer zweifaktoriellen Varianzanalyse werden in Abschnitt 4.3.3 vorgestellt.

Die Darlegungen dieses Abschnitts fassen die Hauptmethode zusammen, die bei der Auswertung aller drei Nutzerstudien, die im Rahmen dieser Arbeit durchgeführt werden, Anwendung findet. Bisher wird jedoch ausschließlich der Einfluss der beiden unabhängigen Variablen betrachtet. Im nächsten Abschnitt wird die Varianzanalyse um die Möglichkeit erweitert, zusätzliche Störvariablen, wie bspw. Vorwissen, Alter oder Geschlecht zu berücksichtigen, was auf das Konzept der Kovarianzanalyse führt.

4.3.2.3. Berücksichtigung von Kovariaten

Die in den vorangegangenen Abschnitten vorgestellten ein- und zweifaktoriellen Varianzanalysen erlauben es ausschließlich, die beobachtete Streuung der Daten in einen Treatment- und einen unspezifizierten Fehleranteil zu zerlegen. Im Rahmen von Laborexperimenten können jedoch häufig weitere bspw. personenbezogene Unterschiede zwischen den Testteilnehmern, wie Vorwissen, Geschlecht oder Alter mit erhoben werden, die ihrerseits zur Varianzaufklärung beitragen können (vgl. Abschn. 4.2.3). Die Erweiterung der Varianzanalyse, um solche Störvariablen oder Kovariaten zu berücksichtigen, führt auf das Konzept der Kovarianzanalyse bzw. Analysis of Covariances (ANCOVA) (Field et al., 2012, S. 462 ff.). Ziel ist es also, die beobachtete Streuung der Messwerte um den Effekt weiterer Einflussgrößen, welche die abhängige Variable beeinflussen könnten, zu korrigieren. Dies dient dazu, die Variation der Messergebnisse innerhalb der Untersuchungsgruppen zu reduzieren, indem der Einfluss dieser Größen herauspartialisiert und damit kontrolliert wird. Im Sinne der Zerlegung der Quadratischen Abweichungen soll das Berücksichtigen der Kovariate also idealerweise dazu führen, dass der um den Wert der Kovariate korrigierte Fehleranteil ($QA_{Fehler,korr}$) geringer ausfällt als der Fehleranteil (QA_{Fehler}) ohne Berücksichtigung der Kovariate. Dies würde im Umkehrschluss die F-Werte der Signifikanztests der Haupteffekte und der Interaktion erhöhen, vorausgesetzt, dass nicht auch die korrigierten Variationsanteile des Treatments geringer ausfallen.

Für eine explizite Darstellung der Rechenschritte zu Durchführung einer Kovarianzanalyse, sowohl im ein- als auch im zweifaktoriellen Fall, sei auf Bortz (2005, S. 362 ff.) verwiesen. Im Folgenden werden hingegen nur die Ideen grob skizziert sowie kurz auf die statistischen Voraussetzungen eingegangen. Im ersten Schritt wird zunächst eine normale ANOVA für die unabhängigen Variablen durchgeführt, ohne dabei die Kovariate zu berücksichtigen um anschließend vergleichen zu können, wie sich die Ergebnisse bei Hinzunahme der Kovariate ändern. Im nächsten Schritt wird ausgehend von der Annahme, dass ein linearer Zusammenhang zwischen der Kovariate und der abhängigen Variable besteht, mit Hilfe einer linearen Rekursion für jede Faktorstufenkombination der beiden unabhängigen Variablen der Einfluss der Kovariate ermittelt. Unter der weiteren Annahme, dass der Einfluss der Kovariate über alle Untersuchungsgruppen hinweg homogen, d.h. konstant ist, wird aus diesen Einzelwerten ein Gesamtwert für den Einfluss der Kovariate auf die abhängige Variable ermittelt. Dieser Gesamtwert wird nun verwendet, um alle Quadratsummen QA_S, QA_E, $QA_{S\times E}$ und QA_{Fehler} um den Einfluss der Kovariate zu korri-

gieren. Um auf signifikante Effekte zu testen, werden anschließend die benötigten F-Tests jeweils mit den korrigierten Quadratsummen durchgeführt.

Zum Abschluss dieses Abschnitts soll noch auf die statistischen Voraussetzungen für die Durchführung einer Kovarianzanalyse eingegangen werden. Diese umfassen zunächst die in Abschnitt 4.3.2.1 bereits erwähnten Voraussetzungen der normalen Varianzanalyse. Hinzu kommt noch die schon angesprochene Homogenität der Regressionssteigungen, die einen konstanten Einfluss der Kovariate auf die abhängige Variable über alle Untersuchungsgruppen hinweg sicherstellt. Dazu ist im Wesentlichen zu testen, ob eine signifikante Wechselwirkung zwischen der Kovariate und den unabhängigen Variablen besteht (Field et al., 2012, S. 468). Ist die Homogenität der Regression nicht gegeben, muss ein erweitertes Datenmodell verwendet werden, das dieser Inhomogenität Rechnung trägt (vgl. Abschn. 4.3.3). Des Weiteren sollte überprüft werden, dass die Treatmenteffekte nur schwach mit der Kovariaten korrelieren (ebd., S. 464 f.). Für personenbezogene Störvariablen kann solch eine Abhängigkeit durch eine randomisierte Zuteilung der Testpersonen zu den einzelnen Untersuchungsgruppen verhindert werden, um bspw. auszuschließen, dass nur Testpersonen einer gewissen Altersgruppe mit dem besseren System arbeiten. In manchen Fällen lässt sich eine gewisse Abhängigkeit zwischen Kovariate und Treatment allerdings nicht vermeiden. So wird im Rahmen des dritten Experiments bspw. der individuelle Sucherfolg als Kovariate bei der Auswertung der Benutzerzufriedenheit verwendet (vgl. Abschn. 7.4.6.2). In diesem Fall ist davon auszugehen, dass die Suchleistung positiv mit der verwendeten Systemqualität korreliert. Bortz (2005, S. 369 f.) merkt jedoch an, dass in solch einem Fall die Berücksichtigung der Kovariaten lediglich den Treatmenteinfluss vermindern kann und somit kein fälschlicherweise signifikanter Zusammenhang angezeigt wird. Bei Verletzung der statistischen Voraussetzungen kann auch im Fall der Kovarianzanalyse auf robuste Verfahren zurückgegriffen werden, die in Abschnitt 4.3.3 behandelt werden.

Zusammenfassend erlaubt die Kovarianzanalyse die Kontrolle von Störvariablen, um bspw. das Vorwissen oder das Alter der Testteilnehmer aus den Messdaten herauszupartialisieren. Um auch dynamische Aspekte des Suchverhaltens analysieren zu können, muss das Konzept der Varianzanalyse jedoch so erweitert werden, dass sie auch die Analyse abhängiger Messdaten, wie sie aus der wiederholten Testung derselben Versuchsperson resultieren, erlaubt. Im Folgenden wird daher die Auswertung gemischter ANOVAs mit Messwiederholung vorgestellt.

4.3.2.4. Messwiederholung und gemischte Modelle

Die dritte Forschungsfrage des hier angestrebten Forschungsvorhabens bezieht sich explizit auf die Dynamik des Nutzerverhaltens und der Relevanzwahrnehmung in Abhängigkeit von Systemgüte und Erwartungshaltung. In diesem Sinne ist es unerlässlich, die Suchleistung und die Zufriedenheit der Testteilnehmer zu unterschiedlichen Zeitpunkten miteinander zu vergleichen. Eine wesentliche Voraussetzung der bisher beschriebenen varianzanalytischen Verfahren ist jedoch gerade, dass die betrachteten Messwerte mit Ausnahme des Treatments keine weiteren systematischen Abhängigkeiten untereinander zeigen. An dieser Stelle setzt die Varianzanalyse mit Messwiederholung an, die es erlaubt Messwerte, die sich aus der wiederholten Beobachtung derselben Testperson ergeben, zu analysieren. Im Gegensatz zu einer normalen einfaktoriellen Varianzanalyse, werden also nicht die Unterschiede zwischen verschiedenen Untersuchungsgruppen, sondern die Änderung der Messwerte innerhalb der Daten einer Testperson betrachtet. Dies

entspricht gerade dem in Abschnitt 4.1.2.1 vorgestellten Within-Subject-Design. Der folgende Abschnitt gibt einen kurzen Überblick über die Grundprinzipien der Methodik sowie die Kombination von Messwiederholungsfaktoren und unabhängigen Variablen, wie sie für die vorliegende Arbeit relevant ist.

Zur Illustration soll zunächst die folgende experimentelle Studie betrachtet werden. Eine Gruppe von 20 Testpersonen bearbeitet mit einer einzelnen Suchmaschine nacheinander drei Suchaufgaben. Die abhängige Variable stellt die jeweils erreichte Suchleistung dar. Somit liegen für jeden Testteilnehmer drei Messwerte vor und es könnte bspw. untersucht werden, ob Lerneffekte zu einer Zunahme der Suchleistung im Laufe der drei Aufgaben führen oder Ermüdungseffekte dominant sind. Da die drei Messwerte pro Teilnehmer nicht unabhängig voneinander sind, wird auf eine Varianzanalyse mit Messwiederholung zurückgegriffen.

Auch die Varianzanalyse mit Messwiederholung folgt dem Grundgedanken, die in der Stichprobe beobachtete Variation der Messwerte in einen Treatment-, hier die Aufgabenposition, und einen Fehleranteil aufzuteilen. Die Summe der quadratischen Abweichungen vom Gesamtmittelwert (QA_{Total}) wird dabei über alle Versuchspersonen und alle Messzeitpunkte gebildet. Zunächst wird QA_{Total} in zwei Anteile aufgespalten (ebd., S. 332): die Variation der Messdaten zwischen den Testpersonen ($QA_{zw\ Vpn}$) und die Variation, welche die Änderung der Messwerte innerhalb der Testpersonen über die drei Aufgaben hinweg beschreibt ($QA_{in\ Vpn}$). Im Gegensatz zur normalen ANOVA sind die Unterschiede zwischen den Testpersonen nicht von Interesse, da sie lediglich individuelle Leistungsunterschiede beschreiben. Stattdessen soll untersucht werden, welchen Einfluss die Aufgabenposition auf die Suchleistung hat. Wie im Fall einer ANOVA ohne Messwiederholung wird $QA_{in\ Vpn}$ nun noch einmal in einen Treatmentanteil ($QA_{Position}$) sowie einen Fehlerterm (QA_{Fehler}) aufgespalten. Der Treatmentanteil beschreibt dabei, wie sich die Messwerte über die Zeit für alle Versuchspersonen ändern. Analog zur Grundgleichung der normalen ANOVA ergibt sich die Relation

$$QA_{Total} = QA_{zw\ Vpn} + QA_{in\ Vpn} \tag{4.4}$$

$$= QA_{zw\ Vpn} + QA_{Position} + QA_{Fehler}\ . \tag{4.5}$$

Um zu testen ob sich in Bezug auf die Aufgabenposition ein signifikanter Effekt feststellen lässt, wird erneut ein F-Test für das Verhältnis von $\sigma^2_{Position}$ und σ^2_{Fehler} durchgeführt.

Sind neben einem Messwiederholungsfaktor noch weitere unabhängige Variablen vorhanden, spricht man von einem *gemischten Design*, da es sich um eine Kombination aus Between-Subjects- und Within-Subject-Versuchsplan handelt (vgl. Abschn. 4.1.2). Im Rahmen dieser Arbeit tritt dieser Fall ein, wenn der Einfluss von Systemgüte und Erwartungshaltung in Abhängigkeit von der Aufgabenposition betrachtet werden soll. Auch in diesem Fall kann eine Varianzanalyse mit Messwiederholung durchgeführt werden. Dabei müssen nun jedoch innerhalb der Variation zwischen den Testpersonen ($QA_{zw\ Vpn}$) auch die Anteile von Systemgüte und Erwartungshaltung identifiziert sowie Wechselwirkungen zwischen zwei oder allen drei unabhängigen Variablen berücksichtigt werden. Die Signifikanz der Haupteffekte und Wechselwirkungen wird wiederum mit Hilfe von F-Tests über einen Vergleich der jeweiligen Effekt- und Fehlervarianz ermittelt.

Die statistischen Voraussetzungen entsprechen in Bezug auf die unabhängigen Variablen ohne Messwiederholung denen der normalen ANOVA (vgl. Abschn. 4.3.2.1). Da jedoch mehrere

Messwerte pro Versuchsperson in die Auswertung eingehen, ist die Voraussetzung der Unabhängigkeit der Fehlerkomponenten verletzt. Im Fall der ANOVA mit Messwiederholung oder für gemischte Designs kann diese jedoch durch die sog. *Sphärizität* ersetzt werden (Field et al., 2012, S. 551), die im Wesentlichen verlangt, dass die Unterschiede zwischen allen Paaren von Untersuchungsgruppen dieselbe Varianz aufweisen müssen. Diese Voraussetzung kann mit Hilfe von Mauchlys Sphärizitätstest überprüft werden. Bei Verletzung dieser Annahme müssen die verwendeten F-Tests entsprechend korrigiert werden (Bortz, 2005, S. 354 ff.).

Nach dieser Übersicht über gemischte Designs, welche die Analyse dynamischer Effekte in Bezug auf Systemgüte und Erwartungshaltung erlauben, wird im Folgenden auf die konkrete Durchführung der Methoden mit der Statistiksoftware R eingegangen.

4.3.3. Umsetzung mit R

Zur numerischen Durchführung der in diesem Abschnitt beschriebenen varianzanalytischen Verfahren wird auf die Statistiksoftware R zurückgegriffen (R Core Team, 2017). Dabei handelt es sich um eine ursprünglich von Ross Ihaka und Robert Gentleman an der Universität Auckland entwickelte und frei unter der GNU-GPL erhältliche Programmiersprache mit einem Schwerpunkt auf statistischen Anwendungen. Die Hauptvorteile des Softwarepakets liegen neben seiner freien Verfügbarkeit besonders in der großen Zahl frei verfügbarer Bibliotheken, die Funktionen zur Durchführung komplexer statistischer Analysen, wie bspw. robuste und nichtparametrische Verfahren, bereitstellen. Darüber hinaus bietet R als Programmiersprache die Möglichkeit Auswertungsschritte in umfangreichen Datenmengen zu automatisieren. Im Rahmen der vorliegenden Arbeit hat dies den Vorteil, dass die Auswertung der unterschiedlichen Benutzerleistungs- und Zufriedenheitsindikatoren weitgehend automatisiert werden kann, sodass nicht jede Analyse einzeln von Hand ausgeführt werden muss. Der folgenden Abschnitt und Tabelle 4.7 geben einen kurzen Überblick über die im Rahmen dieser Arbeit verwendeten Pakete.

4.3.3.1. Varianzanalytische Verfahren

Für alle in Abschnitt 4.3.2.1 diskutieren varianzanalytischen Verfahren existieren R-Pakete, welche die entsprechenden Methoden implementieren und zur Nutzung in der Statistiksoftware R bereitstellen (Field et al., 2012). Tabelle 4.7 fasst die speziell im Rahmen dieser Arbeit gewählten Pakete zusammen. Neben den klassischen parametrischen Verfahren, deren Anwendung an die in Abschnitt 4.3.2.1 diskutierten Voraussetzungen geknüpft sind, stellt R über Bibliotheken auch robuste Testverfahren zur Verfügung, die bei Verletzung dieser Annahmen verwendet werden können. Im Rahmen dieser Arbeit wird dabei überwiegend auf das von Wilcox und Schönbrodt (2014) entwickelte WRS-Paket zurückgegriffen. Dieses implementiert robuste ANOVA-Methoden auf Grundlage von getrimmten Mittelwerten, bei denen ein vorgegebener Anteil extrem vom Mittelwert abweichender Datenpunkte von der Analyse ausgeschlossen wird, was das Verfahren stabil gegenüber Verletzungen der Normalverteilungs- und Varianzhomogenitätsannahme macht (Wilcox, 2011a; Wilcox, 2011b). Über einfache zweifaktorielle Between-Subjects-Designs hinaus, können auch gemischte Designs mit zusätzlichen Within-Subject-Faktoren sowie Kovariaten bei der Auswertung berücksichtigt werden, wodurch alle in Abschnitt 4.3.2.1 beschriebenen Anwendungsfälle abgedeckt sind.

Tab. 4.7.: Übersicht über in dieser Arbeit verwendete R-Pakete. Die Darstellung beschränkt sich auf Bibliotheken, die entsprechende statistische Verfahren implementieren, Pakete zur reinen Datenverarbeitung und Graphikerstellung sind hingegen nicht einzeln aufgeführt.

R-Paket	Beschreibung	Nachweis
stats	Grundlegende von R bereitgestellte Statistikfunktionen. Diese umfassen bspw. den Shapiro-Wilk-Test auf Normalverteilung, den Bartlett-Test auf Sphärizität, die Berechnung von Korrelationsmatrizen für Itemanalysen und Skalenbildung mit Hilfe der Funktion *cor*, sowie das Fitten varianzanalytischer Modelle unter Verwendung der Funktion *aov*.	R Core Team (2017)
psych	Sammlung von Funktionen für die Persönlichkeitsforschung, Psychometrie und Psychologie. Erlaubt mit der Funktion *principal* die Durchführung einer Hauptkomponentenanalyse. Ergänzend werden Funktionen zur Ermittelung der angemessenen Faktorenanzahl mit Hilfe der in Abschnitt 4.3.1 genannten Verfahren, Eigenwertkriterium, Parallelanalyse, VSS-Kriterium, MAP-Kriterium und Scree-Plot bereitgestellt. Des Weiteren erlaubt das Paket die Berechnung von Cronbachs Alpha zur Reliabilitätsanalyse mit Hilfe der Funktion *alpha*.	Revelle (2016)
car	Stellt erweiterte statistische Funktionen bereit. Dazu zählen eine Implementation des Levene-Tests sowie die Berechnung von Typ-II und III ANOVA-Tabellen und Kovarianzanalysen auf Grundlage von aov-Objekten.	Fox und Weisberg (2011)
WRS	Implementierung robuster varianzanalytischer Verfahren auf Grundlage von getrimmten Mittelwerten. Dies umfasst die Funktionen *t1way* und *t2way*, welche die entsprechenden ein- und zweifaktoriellen Verfahren bereitstellen sowie die Funktionen *bbwtrim* und *bbwmcp* für die Berechnung gemischter Designs mit einem Within-Subject- und zwei Between-Subjects-Faktoren wie sie in Experiment 3 für die unabhängigen Variablen Aufgabenposition, Systemleistung und Erwartungshaltung benötigt werden.	Wilcox und Schönbrodt (2014)
npsm	Paket für parameterfreie statistische Verfahren in R. Stellt die Möglichkeit zur Durchführung robuster Kovarianzanalysen bei heterogenen Regressionssteigungen über die Funktion *kancova* zur Verfügung.	Kloke und McKean (2014)
afex	Paket zur Auswertung gemischter Designs, die sowohl Within-Subject- als auch Between-Subjects-Faktoren mit Hilfe klassischer varianzanalytischer Methoden enthalten. Die entsprechende Funktion *aov_ez* kommt im Rahmen von Experiment 3 zum Einsatz, um den Einfluss von Aufgabenposition, Systemgüte und Erwartungshaltung zu untersuchen, falls alle statistischen Voraussetzungen einer klassischen Varianzanalyse erfüllt sind.	Singmann et al. (2016)

4.3.3.2. Faktorenanalyse

Analog zu den varianzanalytischen Verfahren stellt R über Bibliotheken auch faktoranalytische Auswertungsmethoden bereit. Dies umfasst Item-, Hauptkomponenten- sowie Reliabilitätsanalysen. Auch die benötigten Vortests wie Bartlett-Test und KMO-Kriterium können mit Hilfe von Zusatzpaketen direkt in R überprüft werden. Die speziell verwendeten Pakete sind erneut in Tabelle 4.7 aufgeführt.

4.3.3.3. Graphikerstellung

Ein weiterer Vorteil von R besteht in der Möglichkeit, mit geringem Aufwand qualitativ hochwertige Graphiken zu erzeugen, deren Darstellung sich darüber hinaus einfach anpassen lässt. Eine flexibel einsetzbare Bibliothek, die viele unterschiedliche Graphik- und Diagrammtypen bereitstellt ist dabei das von Hadley Wickham entwickelte ggplot2-Paket (Wickham, 2009). Die meisten in dieser Arbeit enthalten Graphen sind mit Hilfe dieses Pakets generiert. Um die in R erstellten Grafiken direkt in Latex nutzbar zu machen, wird auf das Paket tikzDevice zurückgegriffen, das den Export der R-Graphikobjekte als in Latex nutzbaren Tikz-Code erlaubt (Sharpsteen u. Bracken, 2016). Darüber hinaus ermöglicht das Paket Stargazer den direkten Export von R-Tabellen in Latexsyntax (Hlavac, 2015).

4.4. Fazit: Methodisches Vorgehen

Zusammenfassend wird in diesem Kapitel aufbauend auf den in Kapitel 2 und 3 diskutierten Studien ein methodisches Vorgehen für die im Rahmen dieser Arbeit durchgeführten IIR-Benutzerstudien erarbeitet. Neben Fragen der praktischen Testplanung, wie des Untersuchungsdesigns, der Wahl eines Testsystems, der Konstruktion adäquater Testaufgaben sowie dem Aufbau einer geeigneten Testkollektion wird insbesondere die Operationalisierung der unabhängigen Variablen Systemgüte und Erwartungshaltung diskutiert. Dies umfasst die Auswahl des im Rahmen der Untersuchungen realisierten Systemunterschieds sowie die Auswahl einer adäquaten Manipulationsstrategie für die Erwartungshaltung der Testpersonen. In beiden Fällen wird vor dem Hintergrund der Forschungsliteratur ein methodisches Vorgehen für die konkrete Umsetzung im Rahmen der durchgeführten Nutzerstudien entwickelt. Gleiches gilt für die Operationalisierung der erhobenen Messgrößen, die sich in die drei Untergruppen Relevanzwahrnehmung, Benuzterleistung und Benutzerzufriedenheit untergliedern lassen und jeweils unterschiedliche methodische Herangehensweisen erfordern. Im Bezug auf die Relevanzmessung bedeutet dies im Wesentlichen die Auswahl einer geeigneten Bewertungsskala für die Relevanzbewertung durch die Testteilnehmer, wobei die gängige binäre Skala im dritten Experiment auf eine 4- bzw. 8-stufige Relevanzskala erweitert wird. Im Fall der Benutzerleistung kann auf in der IR-Forschung gut etablierte Leistungsindikatoren zurückgegriffen werden, die darüber hinaus im Sinne einer möglichst breiten Erfassung des individuellen Sucherfolgs um weitere Maße insbesondere in Bezug auf die Imprecision erweitert werden. Im Rahmen der Benutzerzufriedenheit existiert hingegen kein standardisiertes Fragebogeninstrument aus dem IR-Bereich, sodass zur Entwicklung geeigneter Frageitems auf das in der Informationssystemforschung weit verbreitete EUCS-Instrument zurückgegriffen wird. Abschließend werden die zur Auswertung der auf diese Weise erhobenen Daten herangezogenen statistischen Verfahren erläutert und Hinweise zur konkreten Umsetzung der Auswertung mit der Statistiksoftware R gegeben. Dabei umfassen die verwendeten Methoden neben faktoranalytischen Verfahren zur Skalenbildung im Kontext der Nutzerzufriedenheit insbesondere unterschiedliche varianzanalytische Verfahren, die es erlauben das gewählte Untersuchungsdesign adäquat mit Hilfe statistischer Modelle abzubilden. Die folgenden drei Kapitel stellen die konkrete Umsetzung der hier erarbeiteten methodischen Vorgehensweisen im Rahmen der durchgeführten experimentellen Studien zum Informationssuchverhalten sowie die jeweils erhaltenen Ergebnisse vor.

5. Experiment 1: Vorstudie zum C/D-Paradigma im IR-Kontext

Aufbauend auf den Ergebnissen von Lamm (2008) werden im Rahmen dieser Arbeit zwei experimentelle Untersuchungen durchgeführt, die den Einfluss von Erwartungen auf die Qualitätswahrnehmung von Suchergebnissen untersuchen. Die folgenden drei Kapitel beschreiben sowohl das methodische Vorgehen als auch die Ergebnisse dieser insgesamt drei Experimente. Die einzelnen Kapitel sind so aufgebaut, dass zunächst das Untersuchungsziel erläutert wird. Anschließend erfolgt die Ableitung von Hypothesen, die im Rahmen der Auswertung mit Hilfe von varianzanalytischen Verfahren überprüft werden. Die Darstellung des Untersuchungsdesigns umfasst die Spezifizierung der unabhängigen und abhängigen Variablen, den Umgang mit Störvariablen, die Beschreibungen von Testkorpus und Testsystem sowie eine Kurzzusammenfassung des Experimentablaufs. Nach einer Diskussion der Pretestergebnisse folgt die Darstellung der Ergebnisse der Hauptuntersuchung. Diese ist in allen drei Fällen in vier Abschnitte gegliedert. Im ersten Abschnitt wird die verwendete Stichprobe beschrieben und überprüft, inwiefern die für die Auswahl der Teilnehmer gesteckten Ziele erreicht sind. Im Anschluss werden die Ergebnisse der abhängigen Variablen dargestellt und diskutiert. Zum Abschluss jedes Kapitels wird ein Fazit für das Experiment gezogen. Im weiteren Verlauf dieses Kapitels folgt nun eine Zusammenfassung des ersten Experiments (ebd.).

5.1. Untersuchungsziel

Im Gegensatz zur systemorientierten Sichtweise betrachtet der benutzerorientierte Ansatz den Vergleich zwischen Anfragen und Dokumenten nicht als Ergebnis, sondern als Ausgangspunkt der Evaluierung. Im Mittelpunkt stehen der Benutzer und seine Bedürfnisse im Umgang mit dem Suchsystem. Neben der objektiven Messung der Retrievalqualität spielt im benutzerorientierten Ansatz auch das subjektive Empfinden der Nutzer eine entscheidende Rolle. Das erste Experiment greift diese Thematik auf und beschäftigt sich mit der Übertragbarkeit objektiv gewonnener Messwerte auf den Anwendungsfall. Bei der Festlegung des Untersuchungsdesigns steht daher die Entwicklung eines interaktiven Testszenarios im Vordergrund, das der Anwendungssituation möglichst ähnlich ist. Diese Vorgehensweise erscheint notwendig, um zu gewährleisten, dass die erhobenen subjektiven Messergebnisse trotz der Standardisierung der Testsituation das tatsächliche Benutzerverhalten widerspiegeln (vgl. Abschn. 4.1.3.1). Im Sinne einer simulierten Arbeitsaufgabe sollen sich die Testpersonen in die Rolle eines Journalisten hineinversetzen und drei kurze Recherchen durchführen (vgl. Abschn. 4.1.3.2). Um ihnen den Eindruck eines möglichst realistischen, lauffähigen Systems zu vermitteln, wird ein Testsystem entwickelt, das die Suche mit einem realen IR-System simuliert. Mit dem Durchlaufen des Suchprozesses sollen die Testpersonen die Suchergebnisse aus einer aktiven Rolle heraus bewerten. Einen weiteren techni-

schen Faktor einer realistischen Testsituation stellt eine intuitiv bedienbare Benutzerschnittstelle dar. Die Gestaltung der Benutzeroberfläche orientiert sich aus diesem Grund am Bedienkonzept gängiger Suchmaschinen, die den meisten Testpersonen vertraut sind. Eine natürliche Bedienung des Testsystems ist auch aus Gründen der Validität der Untersuchung wichtig, damit eventuelle Schwierigkeiten bei der Bedienung nicht zu systematischen Störeffekten führen. Auch auf Ebene der Testaufgaben wird darauf geachtet, Aufgaben auszuwählen, die den Fähigkeiten der Testteilnehmer angepasst sind, sodass alle Testpersonen die gleichen Voraussetzungen haben. Als Anhaltspunkt für die Bearbeitungszeit der Aufgaben wird eine Höchstzeit von zehn Minuten vorgegeben. Um auch hier realistischere Testbedingungen zu schaffen, steht es den Testpersonen jedoch frei, eine Aufgabe vorzeitig abzuschließen. Die Testpersonen erhalten so eine zeitliche Orientierung, ohne den Anspruch, die gesamte Zeit mit der Suche auszufüllen. Diese Vorgehensweise hat zwei entscheidende Vorteile: Durch die Möglichkeit, die Suche vorzeitig zu beenden, wird die Wahrscheinlichkeit einer unmotivierten Suche reduziert (vgl. Abschn. 2.2). Gleichzeitig zeigt sich im Zuge des Tests, dass viele Teilnehmer eine zeitliche Grenze wünschen.

Als Maße für die Retrievalqualität werden subjektive Zufriedenheitsmaße sowie objektive Benutzerleistungsmaße herangezogen. Die theoretische Basis der Vorhersage des Nutzerverhaltens bildet das in der Kundenzufriedenheitsforschung weit verbreitete C/D-Paradigma (vgl. Abschn. 3.3.1.1). Kundenzufriedenheit ist danach das Ergebnis des Vergleichs zwischen erwarteter und wahrgenommener Leistung.

Eine umfassende Beschreibung dieses Experiments findet sich außerdem im Originaltext (Lamm, 2008) sowie in drei weiteren Veröffentlichungen aus dem Jahr 2010 (Lamm et al., 2010a; Lamm et al., 2010b; Werner, 2010).

5.2. Forschungsleitende Hypothesen

Die zentralen Forschungshypothesen dieses Experiments beziehen sich auf die ersten beiden inhaltlichen Forschungsfragen dieser Arbeit (vgl. Abschn. 1.2). Die erste Forschungshypothese betont dabei den Einfluss der Systemgüte und stellt die Kernfrage des benutzerorientierten Evaluierungsansatzes von IR-Systemen dar. Wie bereits in der Einleitung erwähnt, ist die Frage nach der grundsätzlichen Übertragbarkeit klassischer Retrievaltests auf reale Anwendungskontexte Ausgangspunkt für die stärkere Ausrichtung der Evaluierung von IR-Systemen auf die Interaktion der Benutzer. Im Rahmen des ersten Experiments wird diesem Sachverhalt weiter nachgegangen. Zwar ist zum Zeitpunkt der Untersuchung bereits aus einzelnen Studien bekannt, dass von einem gewissen Kompensationsvermögen der Benutzer für Qualitätsunterschiede beim Ranking ausgegangen werden kann, jedoch erscheint der Stand der Forschung in diesem Bereich noch unbefriedigend, sodass zunächst eine direkte Übertragbarkeit unterstellt wird. Es soll also gezeigt werden, dass subjektive Maße zu den gleichen Ergebnissen kommen wie objektive Evaluierungsmaße. Entsprechend lautet die erste Forschungshypothese:

H1: Die Leistung der Benutzer wird durch die Systemgüte gemäß den Annahmen des systemorientierten Ansatzes positiv beeinflusst.

Die zweite Forschungshypothese hebt den Einfluss der Erwartungshaltung auf die Benutzerzufriedenheit heraus. Von besonderem Interesse ist in diesem Zusammenhang die Frage nach

der Übertragbarkeit des C/D-Paradigmas auf den Kontext der Informationssuche. Unter der Annahme, dass der Suchprozess dem Kaufprozess in vielerlei Hinsicht sehr ähnlich ist (vgl. Abschn. 1.2), wird unterstellt, dass die Vorhersagen des C/D-Paradigmas auch auf den Prozess der Informationssuche zutreffend sind. Dementsprechend lautet die zweite Forschungshypothese:

H2: Die Zufriedenheit der Benutzer wird durch ihre Erwartungshaltung und die Systemgüte gemäß den Annahmen des C/D-Paradigmas beeinflusst.

5.3. Methode

Um den Einfluss von Erwartungen sowie die Übertragbarkeit systemorientierter Evaluierungsergebnisse zu untersuchen, wird ein Between-Subjects-Design zugrunde gelegt, das die gleichzeitige Untersuchung der beiden Einflussfaktoren, Systemgüte und Erwartungshaltung, ermöglicht (vgl. Abschn. 4.1.2.2). Hierzu werden die Teilnehmer zufällig einer der vier in Abbildung 5.1 dargestellten Untersuchungsgruppen zugeteilt. Da die Teilnehmer weder ihre Gruppenzugehörigkeit kennen, noch wissen, dass es verschiedene Versuchsgruppen gibt, handelt es sich bei diesem Experiment um einen Blindversuch. Entsprechend dem Versuchsplan erhalten die Teilnehmer in Gruppe 1 zwar die niedrige Erwartungsmanipulation, jedoch enthalten die präsentierten Ergebnislisten vergleichsweise viele relevante Dokumente. Für Teilnehmer in den Gruppen 2 und 3 stimmen Erwartungsmanipulation und Systemleistung überein. In Gruppe 4 hingegen trifft eine hohe Erwartungshaltung auf eine schlechte Systemleistung. Alle Teilnehmer verwenden das gleiche in Abschnitt 5.3.5 beschriebene Wizard-of-Oz-System zur Bearbeitung der drei Suchaufgaben (A1 bis A3), d.h. den Testpersonen ist nicht bekannt, dass die Reaktionen des Systems vorherbestimmt sind. Zur individuellen Beurteilung der erfahrenen Retrievalqualität werden sowohl subjektive Zufriedenheitsmaße als auch objektive Benutzerleistungsmaße herangezogen. Dabei wird die Zufriedenheit der Probanden einmal im Anschluss an die Bearbeitung aller drei Aufgaben erhoben. Die Messwerte zur Beurteilung der Benutzerleistung werden über die drei Aufgaben gemittelt. Die Mittelwertbildung hat den Vorteil, dass der Einfluss von Ausreißern bei den Einzelmessungen abgemildert wird. Die angestrebte Stichprobengröße beträgt 80 Probanden, welche sich zu gleichen Teilen auf die vier Versuchsgruppen verteilen.

Im Einklang mit dem C/D-Paradigma wird von drei Grundannahmen ausgegangen: 1. Probanden mit unrealistisch niedrigen Erwartungen erleben im Verlauf der Suche eine positive Diskonfirmation und sind infolgedessen mit dem Testsystem zufrieden. 2. Probanden mit realistischerweise niedrigen oder hohen Erwartungen fühlen sich während der Suche in ihren Erwartungen bestätigt und sind deshalb ebenfalls mit dem Testsystem zufrieden. 3. Probanden mit unrealistisch hohen Erwartungen hingegen erleben im Laufe der Suche eine negative Diskonfirmation und sind infolgedessen mit dem Testsystem unzufrieden.

5.3.1. Manipulation der unabhängigen Variablen

In allen drei Experimenten stellen die Systemqualität und die Benutzererwartungen die beiden unabhängigen Variablen des Versuchsaufbaus dar. Im Folgenden wird ihre Operationalisierung im Rahmen des ersten Experiments beschrieben. Um die Systemleistung zu kontrollieren, werden wie in vergleichbaren Studien (vgl. Abschn. 4.2.1.1) künstlich erzeugte Ergebnislisten eingesetzt. Für jede der drei Suchaufgaben (vgl. Abschn. 5.3.4) werden zwei unterschiedliche Ergebnislisten

		System	
		gut	schlecht
Erwartung	niedrig	Gruppe 1 $\frac{A_1+A_2+A_3}{3}$	Gruppe 2 $\frac{A_1+A_2+A_3}{3}$
	hoch	Gruppe 3 $\frac{A_1+A_2+A_3}{3}$	Gruppe 4 $\frac{A_1+A_2+A_3}{3}$

Abb. 5.1.: Versuchsplan des ersten Experiments. In dem gewählten Between-Subjects-Design führt die zweifache Abstufung der beiden Faktoren Systemgüte und Erwartungshaltung zu den dargestellten vier Untersuchungsgruppen. Während die Erhebung der Benutzerzufriedenheit nach Abschluss aller drei Aufgaben erfolgt, wird die erreichte Benutzerleistung über alle drei Suchaufgaben gemittelt.

erzeugt: eine, um eine niedrige und eine, um eine hohe Retrievalqualität zu simulieren. Insgesamt werden also für das erste Experiment sechs verschiedene Listen generiert. Konkret sind die Ergebnislisten für das schlechtere System durch eine Precision von $0{,}5$ und eine AvP von $0{,}55$ gekennzeichnet. Die Ergebnislisten für das bessere System sind hingegen durch eine Precision von $0{,}6$ und eine AvP von $0{,}75$ charakterisiert. Zu ihrer Erstellung wird auf einen in Turpin und Scholer (2006, S. 14) beschriebenen Algorithmus zurückgegriffen. Der Algorithmus generiert jede Ergebnisliste als einen Binärstring. Dabei stehen die Nullen und Einsen jeweils für die Position der irrelevanten bzw. relevanten Dokumente. Da die AvP die gesamte Ergebnisliste berücksichtigt, ist zu beachten, dass die Verteilung irrelevanter und relevanter Dokumente auch bei gleicher AvP sehr unterschiedlich ausfallen kann. So ist es möglich, dass bei einer niedrigen AvP mehr relevante Dokumente auf den ersten zehn Listenplätzen stehen als bei einer hohen. Anders ausgedrückt, kann sich die AvP stark von der entsprechenden Precision der ersten zehn Dokumente unterscheiden (vgl. Abschn. 4.2.1.1). Auf diesen Umstand weisen auch Scholer und Turpin (2009) im Zusammenhang mit der MAP hin. Um trotzdem eine konsistente Manipulation der Systemleistung in Bezug auf die ersten Listenplätze zu gewährleisten, wird deshalb zusätzlich die Precision der ersten fünf Dokumente kontrolliert und nur solche Ergebnislisten gewählt, die auf den ersten fünf Listenplätzen bei niedriger Systemleistung drei, bei hoher Systemleistung hingegen nur ein irrelevantes Dokument enthalten. Die konkreten Listen können Anhang A.1 entnommen werden.

Auch die Erwartung wird in einer niedrigen und einer hohen Ausprägung variiert. Dazu werden sowohl explizite als auch implizite Beurteilungsmerkmale eingesetzt. Zunächst wird allen Testpersonen mitgeteilt, dass die Bibliothek der Universität Hildesheim über den Einsatz einer neuen Suchmaschine für Fachzeitschriften nachdenkt, die nun erstmals in einem Benutzertest evaluiert werden soll. Als expliziten Hinweisreiz erhalten die Teilnehmer je nach Versuchsgruppe unterschiedliche Instruktionen. Während das Testsystem den Gruppen mit hohen Erwartungen als professionelles Suchsystem mit einem Kaufpreis von 20.000 € vorgestellt wird, erhalten die

Gruppen mit niedrigen Erwartungen die Information, dass es sich bei dem System um ein Studentenprojekt einer anderen Hochschule handelt, das im Rahmen eines Projektseminars an der Universität Hildesheim weiterentwickelt werden soll. Während also im ersten Fall der Eindruck vermittelt wird, mit einem qualitativ sehr hochwertigen System zu arbeiten, wird im zweiten Fall die Erwartungshaltung aufgebaut, dass es sich um einen noch in der Entwicklung befindlichen Prototypen handelt. Als impliziter Hinweisreiz wird die Testinstruktion auf qualitativ unterschiedliche Papierarten gedruckt: Zur Unterstützung der niedrigen Erwartungsmanipulation auf mattem und zur Unterstützung der hohen Erwartungsmanipulation auf glänzendem Papier. Der genaue Wortlaut beider Testinstruktionen kann in Anhang A.2 nachgelesen werden.

5.3.2. Operationalisierung der abhängigen Variablen

Zur Beurteilung der wahrgenommenen Retrievalqualität werden subjektive Zufriedenheitsmaße sowie objektive Benutzerleistungsmaße herangezogen. Wie bereits erwähnt, ist die benutzerorientierte Evaluierung zur Zeit des ersten Experiments ein vergleichsweise neuer Ansatz im Bereich des IR. Nur vereinzelt finden sich daher Ansätze, die die Retrievalqualität nicht auf eine rein verhaltensorientierte Weise zu erfassen versuchen, sondern auch die subjektive Zufriedenheit der Benutzer in den Blick nehmen (vgl. Abschn. 3.3). Auch sind die verwendeten Methoden oft noch wenig standardisiert (vgl. Abschn. 4.2.2.3). Im ersten Experiment wird die Zufriedenheit der Testpersonen aus den genannten Gründen anhand eines eigens entwickelten Fragebogens erhoben. Dieser umfasst 15 geschlossene Items, die sich auf unterschiedliche Aspekte der Zufriedenheit der Benutzer mit der Suchmaschine beziehen (vgl. Abschn. 3.3). Als weitere Orientierung dienen von Kaczmirek (2003) und Gediga et al. (2005) entwickelte Fragebögen zur Nutzerzufriedenheit. Eingeschätzt wird die Zufriedenheit mit der Suchmaschine allgemein, mit der Bedienbarkeit, mit der eigenen Suchleistung, mit der Qualität der Ergebnisse, mit dem Ranking sowie mit Recall und Precision. Neben Items, die die Zufriedenheit der Testpersonen in Bezug auf einen dieser Aspekte direkt erfassen, enthält der Fragebogen auch indirekte Items, die typische Indikatoren für Zufriedenheit erheben. Als Beispiel ist Item 7 (*Ich würde Periodikum jederzeit wieder als Suchmaschine verwenden*) zu nennen. Dabei ist Periodikum der Name der zu bewertenden Suchmaschine. Weitere Items dieser Art betreffen die Empfehlung der getesteten Suchmaschine für die Bibliothek der Universität Hildesheim (Item 15) sowie die Bereitschaft an einem weiteren Benutzertest teilzunehmen. Die direkten Items werden mit Angaben auf einer 7-stufigen Bewertungsskala von *trifft vollkommen zu* bis *trifft überhaupt nicht zu* beantwortet, sodass den Befragten je drei Abstufungen hinsichtlich Zustimmung oder Ablehnung sowie eine neutrale mittlere Antwortkategorie zur Verfügung stehen. In Anhang A.3 findet sich ein Überblick über alle verwendeten Frageitems. Eine ausführlichere Beschreibung des Fragebogens findet sich außerdem in Lamm (2008, S. 81 ff.).

Die Suchleistung der Testpersonen wird anhand von fünf Benutzerleistungsmaßen beurteilt, die in Tabelle 5.1 zusammengefasst sind (vgl. Abschn. 4.2.2.2). Die Mehrheit der im ersten Experiment verwendeten Leistungsmaße ist dabei aus anderen Studien übernommen und an das Untersuchungsdesign angepasst. Dies ist vorteilhaft, weil diese bereits erprobt sind. Außerdem wird hierdurch auch die Vergleichbarkeit der Ergebnisse mit anderen Studien erleichtert. Die verwendeten Leistungsmaße bewerten die Vollständigkeit und die Genauigkeit der Suchergebnisse sowie die Zeit, die benötigt wird, um das erste relevante Dokument zu finden.

Tab. 5.1.: Übersicht über verwendete Benutzerleistungsmaße.

Maß	**Beschreibung**	**Quelle**
Anz. richtig rel. Dok. (RRD)	Anz. richtig rel. bew. Dok.	Turpin und Scholer (2006)
Benutzerrecall (BR)	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	Al-Maskari et al. (2006)
Benutzerprecision (BP)	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	Al-Maskari et al. (2006)
Zeit erstes richtig rel. Dok. (T1)	Zeit zum Auffinden des ersten richtig rel. bew. Dok.	Turpin und Scholer (2006)
Pre-Click-Precision(PCP)	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	Resnick und Lergier (2003)

Das erste Leistungsmaß in Tabelle 5.1 ist die Anzahl der richtig als relevant erkannten Dokumente (RRD), die die Testpersonen innerhalb der vorgegebenen Bearbeitungszeit von maximal zehn Minuten finden können. Es ist zu beachten, dass sich diese Menge auf die Dokumente bezieht, die in Übereinstimmung mit den Juroren als relevant bewertet werden. Dieses Leistungsmaß wird auch von Turpin und Scholer (2006) verwendet. Während der Benutzerrecall (BR) ein Maß für die Vollständigkeit des Rechercheergebnisses ist, misst die Benutzerprecision (BP) die Genauigkeit, mit der die Testpersonen relevante Dokumente finden und irrelevante Dokumente zurückweisen. Dazu wird RRD im ersten Fall durch die Gesamtanzahl der relevanten Dokumente in der Ergebnisliste geteilt, im zweiten Fall hingegen durch die Gesamtanzahl der durch die Testperson als relevant bewerteten Dokumente. Beide Benutzerleistungsmaße werden in ähnlicher Form bereits von Al-Maskari et al. (2006) verwendet. In Anlehnung an Turpin und Scholer (2006) wird außerdem die Zeit verglichen, die die Testpersonen benötigen, um ihr erstes richtig als relevant bewertetes Dokument zu finden (T1). Die Pre-Click-Precision (PCP) schließlich bezieht den ersten Eindruck, den die Testpersonen von der Ergebnisliste haben, in die Berechnung der Benutzerleistung mit ein, indem diesmal RRD durch die Gesamtanzahl der durch die Testpersonen als möglicherweise relevant ausgewählten Dokumente geteilt wird. Dieses Leistungsmaß bezieht sich auf eine Studie von Resnick und Lergier (2003), in der die Teilnehmer nach ihrem Vertrauen in die eigene Dokumentenauswahl befragt werden. In diesem Kontext führen Resnick und Lergier (ebd.) den Begriff der *pre-click confidence* ein.

5.3.3. Umgang mit Störvariablen

Die Ausführungen in den Kapiteln 2 und 3 zum Stand der Forschung zeigen, dass der benutzerorientierte Ansatz erheblich komplexer ist als rein systembasierte Retrievaltests. Dies liegt vor allen Dingen darin begründet, dass das individuelle Verhalten und Empfinden der Benutzer nicht ausschließlich durch die untersuchten Variablen bestimmt wird. Vielmehr können die abhängigen Variablen durch weitere Faktoren beeinflusst werden. Aus diesem Grund ist es wichtig die interne Validität der Untersuchung sicherzustellen. Dazu muss der Einfluss möglicher Störfaktoren auf die abhängigen Variablen entweder unterdrückt oder kontrolliert werden. Im Rahmen des hier beschriebenen Experiments werden basierend auf den Ergebnissen und Erfahrungen der im Stand der Forschung diskutierten Studien die im Folgenden beschriebenen Einflussgrößen als mögliche Störvariablen identifiziert.

Um Alterseffekte auszuschließen, wird die Teilnahme auf eine feste Altersgruppe zwischen 18

und 30 Jahren begrenzt. Genauso wie im Fall des Alters, wird auch der Einfluss des Geschlechts durch Konstanthaltung kontrolliert, indem ausschließlich weibliche Testpersonen rekrutiert werden. Um nicht von vornherein Personen nichtdeutscher Muttersprache als Teilnehmer ausschließen zu müssen, wird ein möglicher Effekt der Muttersprache im Zuge der Überprüfung der Daten untersucht. Auch ein möglicher Einfluss der Sucherfahrung der Testpersonen auf die abhängigen Variablen wird bei der Datenüberprüfung berücksichtigt. Dazu wird die Sucherfahrung mit Hilfe von fünf (halb-) offenen Items erfasst (vgl. Anh. A.4). Etwaige Lern- und Reihenfolgeeffekte werden durch Variation der Aufgabenreihenfolge innerhalb der vier Versuchsgruppen minimiert bzw. ausgeschlossen. Um weiterhin sicherzustellen, dass alle Teilnehmer gleiche Ausgangsvoraussetzungen für die Relevanzbewertung haben, werden den Testpersonen, ähnlich wie in Studien von Kaczmirek (2003), Kwahk und Oh (2009) und Huffman und Hochster (2007), die zu verwendenden Suchbegriffe vorgegeben. Zwar bewirkt diese Einschränkung, dass sich die Probanden der Testsituation erneut bewusst werden, doch dient die Konstanthaltung der Suchbegriffe dazu, über soziodemografische Aspekte hinausgehende subjektive Einflüsse bei der Relevanzbewertung auszuschließen. Wie der letzte Punkt bereits zeigt, geht es bei dem Umgang mit möglichen Störvariablen auch darum, zwischen notwendiger Standardisierung auf der einen und angemessenem Realitätsgrad auf der anderen Seite abzuwägen. In Bezug auf die Bearbeitungsdauer der einzelnen Testaufgaben wird deshalb lediglich eine maximale Bearbeitungszeit von zehn Minuten pro Aufgabe vorgegeben. Es steht den Teilnehmern jedoch frei, ihre Suche vorzeitig zu beenden, wenn sie nach eigenem Ermessen genügend relevante Dokumente gefunden haben. Zum einen wird durch diese Vorgehensweise die Wahrscheinlichkeit einer die interne Validität der Untersuchung gefährdenden unmotivierten Suche reduziert. Gleichzeitig wird der Tatsache Rechnung getragen, dass Testpersonen durchaus eine zeitliche Grenze wünschen (vgl. Abschn. 5.3.7). Im Falle einer Konfundierung der Qualitätswahrnehmung mit den oben genannten Störvariablen werden diese Faktoren in den weiteren Analysen als Kovariaten berücksichtigt.

5.3.4. Auswahl des Testkorpus

Im Folgenden werden die Aufgaben, die die Testpersonen zu bewältigen haben, erläutert. Als Testkorpus werden im ersten Experiment Volltexte aus dem deutschsprachigen CLEF-Korpus verwendet. Konkret handelt es sich um drei Suchaufgaben aus den Jahren 2001 und 2003 (Braschler, 2002; Braschler, 2004). Die Nutzung eines bestehenden Testkorpus hat zunächst den ganz pragmatischen Vorteil der Zeitersparnis, weil Suchthemen, Dokumente und Relevanzurteile bereits vorliegen. Ein weiterer Vorteil besteht darin, dass die Materialien bereits erprobt sind. Auch der Vergleich unterschiedlicher Untersuchungen wird auf diese Weise erleichtert. Bei der Auswahl der Aufgaben ist es wichtig, gemeinsame, lebensweltlich bekannte Themen zu wählen, um für alle Testpersonen annähernd gleiche Ausgangsvoraussetzungen zu schaffen (vgl. Abschn. 4.1.3.2). Weitere Kriterien sind ein ähnlicher Schwierigkeitsgrad der Aufgaben untereinander, eine gute Verständlichkeit, keine Über-, aber auch keine Unterforderung sowie eine objektive Auswertbarkeit. Konkret lauten die anhand dieser Kriterien ausgewählten Suchthemen wie folgt:

1. Für einen Beitrag über erneuerbare Energien suchst Du nach Presseartikeln, die die Nutzung von umweltfreundlicher Energie oder eine darauf ausgerichtete Politik betreffen, d.h. von Energie, die aus erneuerbaren Energiequellen erzeugt wurde.

Tab. 5.2.: Beschreibung des verwendeten Testkorpus. Angegeben sind die pro Suchaufgabe vorhandenen relevanten bzw. irrelevanten Dokumente sowie die zur Erstellung der Ergebnislisten benötigten Dokumentenanzahlen.

		Aufteilung	Atomtransporte in Deutschland	Erneuerbare Energien	Kinderarbeit in Asien
Verfügbare	relevant		57	60	50
Dokumente	irrelevant		48	50	42
	gesamt		105	110	92
schlechtes	relevant	50%	48	50	42
System	irrelevant	50%	48	50	42
AvP $= \mathbf{0{,}55}$	gesamt		96	100	84
gutes	relevant	60%	57	60	50
System	irrelevant	40%	39	40	34
AvP $= \mathbf{0{,}75}$	gesamt		96	100	84

2. Für einen Beitrag über Atomtransporte in Deutschland suchst Du Berichte über Proteste gegen den Transport von radioaktivem Müll in Castor-Behältern in Deutschland.

3. Für einen Beitrag über Kinderarbeit in Asien suchst Du nach Presseartikeln, die Kinderarbeit in Asien diskutieren und Vorschläge zu deren Beseitigung oder zur Verbesserung der Arbeitsbedingungen für Kinder liefern.

Eine Übersicht über die Anzahl der verfügbaren Dokumente für die einzelnen Aufgaben und ihre Aufteilung auf die beiden Systemvarianten kann in Tabelle 5.2 eingesehen werden.

5.3.5. Beschreibung des Testsystems

Dieser Abschnitt beschreibt die grundlegenden Eigenschaften des im ersten Experiment verwendeten Testsystems. Eine ausführlichere Darstellung findet sich außerdem in Lamm (2008, S. 63). Um eine möglichst realitätsnahe Testsituation zu schaffen, wird für das erste Experiment ein in Java geschriebenes Anwendungssystem entwickelt, das den Suchprozess einer realen Suchmaschine simuliert. Es wird davon ausgegangen, dass ein solches Wizard-of-Oz-System den Testpersonen dabei hilft, sich in die vorgegebene Nutzungssituation hineinzuversetzen. So sind die Testpersonen nicht mehr nur Juroren, sondern sie bewerten die Dokumente aus einer aktiven Rolle heraus. Dabei stellt die intuitive Bedienbarkeit des Systems einen wichtigen Faktor einer realistischen Testsituation dar. Gestaltung und Funktionalität der Benutzeroberfläche orientieren sich deshalb am Bedienkonzept gängiger Suchmaschinen, die den meisten Testpersonen vertraut sind. Eine intuitive Bedienung ist nicht nur hinsichtlich der Realitätsnähe wichtig, sondern auch aus Gründen der Validität. Ein Teilnehmer, der das Bedienkonzept nicht richtig verstanden hat, wird auch kein valides Ergebnis abliefern. In Anlehnung an die Darstellung gängiger Suchmaschinen werden in der Ergebnisliste zehn Treffer pro Seite angezeigt (vgl. Abb. 5.2). Auch die Trefferbeschreibungen mit Titel, Snippet und Quelle entsprechen von ihrem Erscheinungsbild her typischen Darstellungen von Suchergebnislisten. Da in dem verwendeten CLEF-Korpus keine Snippets vorhanden sind, müssen diese zunächst definiert werden. In Anlehnung an Kaczmirek (2003), wird hier der erste Satz des entsprechenden Volltextes als Snippet angezeigt. Die Relevanzbewertung der Dokumente erfolgt im Volltextfenster, das sich öffnet, sobald einer der Links in der Ergebnisliste angeklickt wird (vgl. Abb. 5.3).

Da das Testsystem in vierfacher Ausführung vorliegt (je eine Version pro Versuchsgruppe), han-

delt es sich bei dem ersten Experiment um einen Blindversuch, bei dem zwar die Testpersonen vom genauen Versuchsaufbau nur so weit informiert werden, als es für die Durchführung des Tests notwendig ist, dem Testleiter jedoch die Gruppenzugehörigkeit bekannt ist (vgl. Abschn. 5.3). Rein äußerlich unterscheiden sich die vier Systemversionen nur durch das Copyright, das angibt, welche Erwartungshaltung im aktuellen Fall manipuliert wird. Weitere Eigenschaften des Testsystems, die hier kurz Erwähnung finden sollen, betreffen zwei Mechanismen, die einen reibungslosen Ablauf des Tests garantieren sollen. Zum einen verhält sich das Testsystem tolerant gegenüber der Reihenfolge der vorgegebenen Suchbegriffe. Um Rechtschreibfehler abzufangen, werden auch Eingaben akzeptiert, die bis zu einer Levenshtein-Distanz von sieben mit den vorgegebenen Suchbegriffen übereinstimmen. Um außerdem ein versehentliches Schließen des Testsystems durch die Testpersonen zu verhindern, wird der Schließbutton während der Testdurchführung gesperrt. Diese Maßnahme ist insbesondere deshalb wichtig, weil die Dokumente jeweils in einem neuen Fenster geöffnet und nach der Bewertung durch die Testpersonen wieder geschlossen werden.

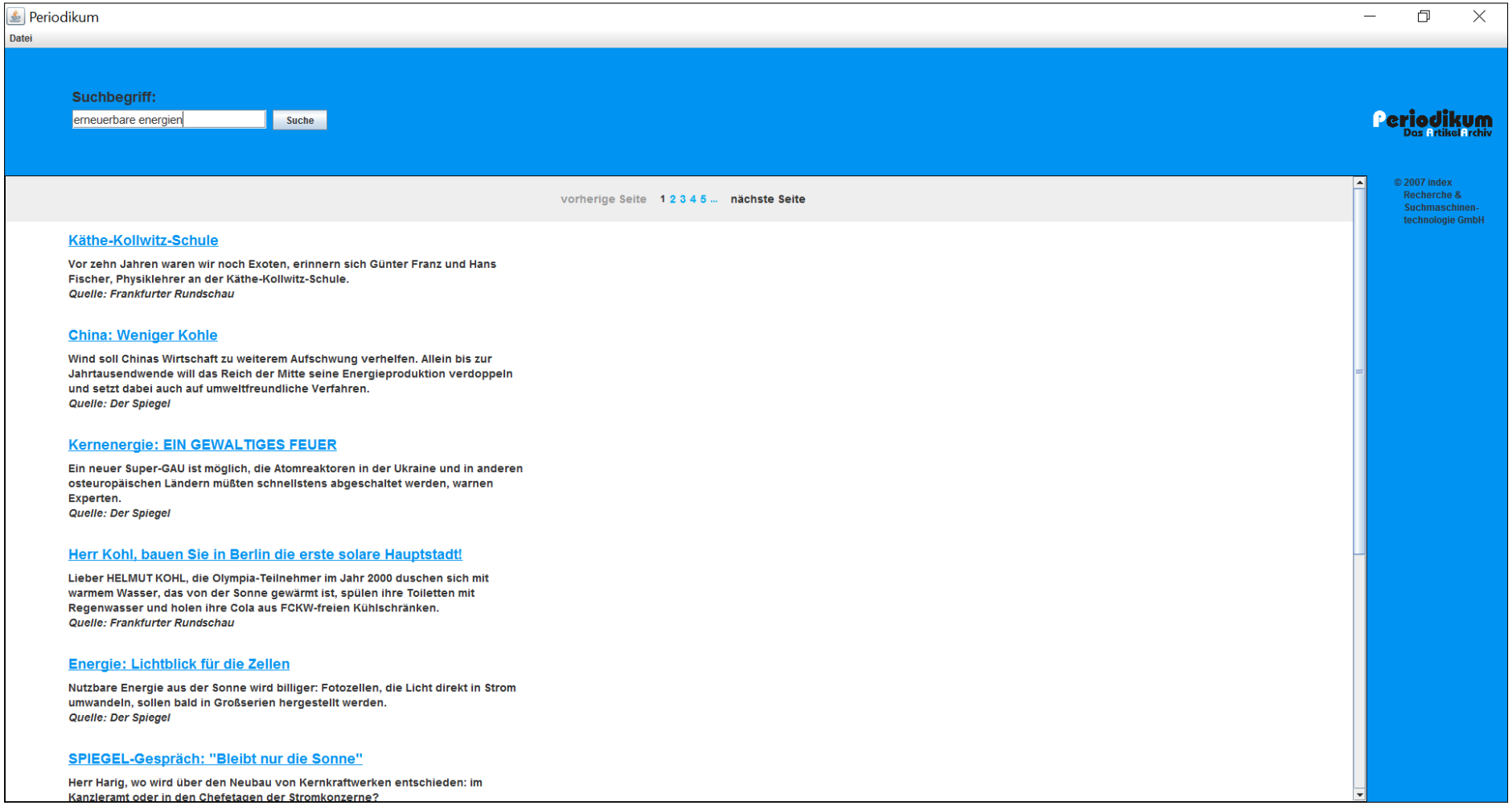


Abb. 5.2.: Testsystem des ersten Experiments: Darstellung der Suchergebnisliste.

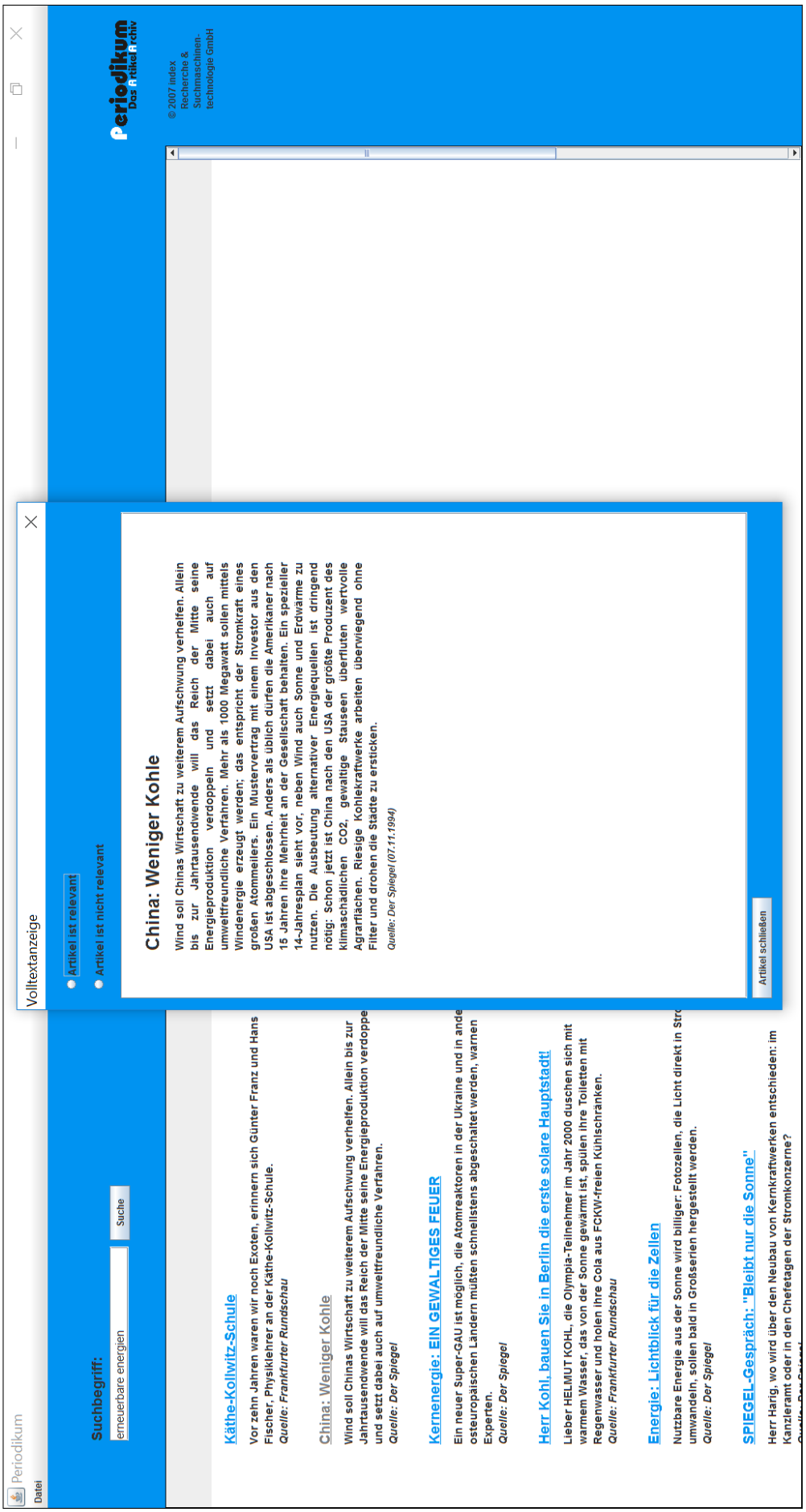


Abb. 5.3.: Testsystem des ersten Experiments: Relevanzbewertung.

5.3.6. Ablauf

In diesem Abschnitt wird der zeitliche Ablauf des Experiments erläutert (vgl. Abb. 5.4). Alle Versuchspersonen werden individuell in einem ruhigen Raum getestet. Der Ablauf ist für alle Testpersonen identisch. Zu Beginn wird ihnen die in schriftlicher Form verfasste Testinstruktion ausgehändigt, in der die Erwartungsmanipulation und die Aufgabe beschrieben werden. Auf Nachfrage werden diese Informationen nochmals mündlich erläutert. Die Testpersonen werden nun gebeten, nacheinander drei Themen mit dem Ziel zu recherchieren, später als Journalist einen Beitrag über diese Themen schreiben zu können. Dabei wird die Reihenfolge der Aufgaben systematisch variiert (vgl. Abschn. 5.3.3), um Lerneffekten vorzubeugen. Je nach Versuchsgruppe arbeiten die Probanden mit einer der vier in Abschnitt 5.3.5 beschriebenen Systemvarianten. Beim Suchen mit den vorgegebenen Suchbegriffen (vgl. Abschn. 5.3.3) wird darin die der Versuchsgruppe zugeordnete Trefferliste angezeigt, die bis zur Beendigung der Aufgabe beibehalten wird. Die Probanden können selbst entscheiden, welche und wie viele Treffer sie sich genauer ansehen möchten. Wird jedoch ein Dokument ausgewählt, muss dieses bewertet werden, bevor weitere Treffer aufgerufen werden können. Nach Bearbeitung aller drei Aufgaben werden die Testpersonen gebeten, ihre Eindrücke mit Hilfe des in Abschnitt 5.3.2 beschriebenen Fragebogens festzuhalten. Neben den dort genannten Items zur Erfassung der Benutzerzufriedenheit enthält dieser Fragebogen auch die zur Auswertung erforderlichen Fragen zur Person (Alter, Geschlecht, Muttersprache etc.). Darüber hinaus werden die Teilnehmer in einer offenen Frage nach weiteren als wichtig erachteten Aspekten gefragt, die entweder die Suchmaschine oder das Experiment betreffen. Als Anreiz und Belohnung für die geopferte Zeit haben alle Teilnehmer am Ende der Untersuchung die Gelegenheit, an der Verlosung von drei Geldpreisen teilzunehmen. Die Gesamtdauer des Experiments liegt bei etwa 45 Minuten.

5.3.7. Ergebnisse des Pretests

Im Vorfeld der eigentlichen Untersuchung wird das gesamte experimentelle Material einem Pretest mit vier Versuchspersonen unterzogen. Somit ist jede Versuchsgruppe einmal vertreten. Mit der qualitativen Voruntersuchung werden in erster Linie die Funktionalität des Testsystems und die Verständlichkeit des Fragebogens überprüft. Aber auch weitere Fragen, wie die Qualität der Instruktion und der zeitliche Ablauf sollen im Rahmen des Pretests überprüft und ggf. überarbeitet werden. Um Zugang zu den Verstehensprozessen der Versuchspersonen zu finden, wird die Methode des lauten Denkens gewählt. Ergänzend werden die Versuchspersonen während der Aufgabenbearbeitung beobachtet, um über ihre subjektiven Äußerungen hinaus einen Eindruck von der Qualität des Untersuchungsmaterials zu erhalten. Die Ergebnisse des Pretests führen zu einigen grundlegenden Änderungen und Anpassungen, von welchen die wichtigsten im Folgenden kurz erläutert werden. Eine ausführliche Darstellung der Pretestergebnisse findet sich außerdem in Lamm (2008, S. 71).

In Bezug auf die Funktionalität des Testsystems stellt sich heraus, dass eine einzelne lange Trefferliste ohne Seitenunterteilungen nicht den gewohnten Erwartungen an typische Trefferdarstellungen entspricht, sodass die einzelnen Trefferbeschreibungen für die Hauptuntersuchung auf mehrere Seiten verteilt werden. Der Fragebogen erweist sich als verständlich und zeitlich ausreichend kalkuliert, sodass hier keine Änderungen notwendig sind. Durch den Pretest können weiterhin zwei Schwachstellen bezüglich der Instruktion der Testpersonen behoben werden. Zum

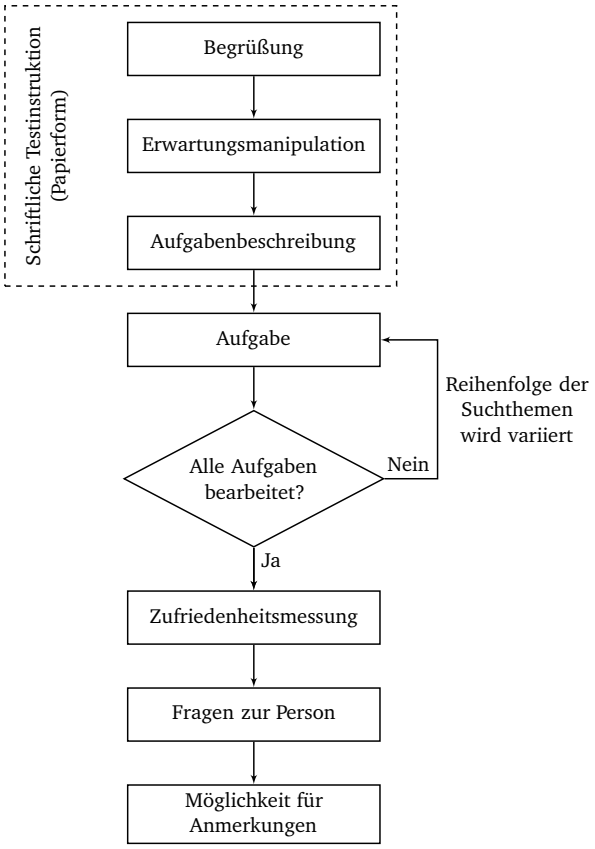


Abb. 5.4.: Schematische Darstellung des Versuchsablaufs des ersten Experiments.

einen ist allen vier Versuchspersonen nicht klar, ob sie die aufgerufenen Dokumente tatsächlich nach ihrer Relevanz bewerten sollen. Zum anderen stellt die mangelnde Aktualität der Dokumente eine Unsicherheit für die Versuchspersonen dar. In beiden Fällen wird die Instruktion entsprechend geändert. Eine weitere Anpassung betrifft die Bearbeitungszeit der Aufgaben. Die ursprüngliche Idee, die Probanden das Ende der Suche selbst bestimmen zu lassen, stellt sich im Pretest als unzureichend heraus, da diese völlige Offenheit zu einer spürbaren Verunsicherung der Teilnehmer führt. Aus diesem Grund wird in der Hauptuntersuchung eine maximale Bearbeitungsdauer von zehn Minuten vorgegeben. Die letzte Anpassung bezieht sich auf die Güte der Ergebnislisten. Die Tatsache, dass alle Versuchspersonen der Voruntersuchung angeben, dass die Ergebnislisten nicht gut gefiltert sind, führt dazu, dass die Precision der besseren Ergebnislisten in der Hauptuntersuchung von $0{,}5$ auf $0{,}6$ angehoben wird.

5.4. Ergebnisse

Die Hauptuntersuchung wird im Frühjahr 2008 an der Universität Hildesheim durchgeführt. Dieser Teil des Kapitels stellt die Ergebnisse des ersten Experiments dar. Dazu wird zunächst die Repräsentativität der Stichprobe diskutiert. Im Ergebnisteil wird die Suchleistung der Testpersonen anhand der fünf in Abschnitt 5.3.2 vorgestellten Leistungsmaße untersucht. Weitere Ergebnisse umfassen die subjektiven Zufriedenheitsurteile der Testpersonen sowie die Änderungen beider Variablengruppen in Abhängigkeit von den in Abschnitt 5.3.1 vorgestellten Ausprägungen der unabhängigen Variablen. Eine eingehende Überprüfung der Gütekriterien des Experiments soll außerdem klären, ob es außer den beiden unabhängigen Variablen noch weitere Einflussgrößen gibt, deren Einbeziehung sinnvoll und notwendig sein kann.

5.4.1. Beschreibung der Stichprobe

Die Probanden des ersten Experiments rekrutieren sich großenteils aus Studierenden der Universität Hildesheim. Die Stichprobe umfasst 89 weibliche Testpersonen, deren demographische Merkmale Tabelle 5.3 zu entnehmen sind. Da zu Beginn der Hauptuntersuchung nicht abzusehen ist, ob die angestrebte Stichprobengröße von 80 Probanden erreicht wird, werden zunächst auch Testpersonen zugelassen, die etwas jünger oder älter als 18 bzw. 30 Jahre sind. Das Alter der Teilnehmer reicht dementsprechend von 17 bis 32 Jahren, das Medianalter beträgt 24 Jahre. Diese geringe Abweichung von den ursprünglich für die Auswahl der Teilnehmer gesteckten Zielen erscheint jedoch vor dem Hintergrund einer geplanten Überprüfung von Alterseffekten durchaus vertretbar. Die Muttersprache von 13,48% der Testpersonen ist nicht deutsch, was einer absoluten Zahl von 12 Teilnehmern entspricht. Hinsichtlich der Tätigkeit der Testpersonen wird zwischen den Kategorien Schülerin, Auszubildende, Studierende, Berufstätige und Sonstige unterschieden, woraus sich ein prozentualer Anteil von 78,65% Studierender gegenüber Testpersonen anderer Tätigkeiten ergibt.

Tab. 5.3.: Demographische Daten ($n = 89$).

Variable	**Maß**	**Wert**	
Alter	Median	24	
	Standardabweichung	3,21	
	Spanne	17 – 32	
	Mittelwert	24,03	
	Kategorie	**Anzahl**	**Prozent**
Muttersprache	Deutsch	77	86,52
	nicht Deutsch	12	13,48
Tätigkeit	Schülerinnen	7	7,87
	Auszubildende	2	2,25
	Studierende	70	78,65
	Berufstätige	5	5,62
	Sonstige	5	5,62

Tabelle 5.4 fasst die Ergebnisse für die erhobenen Items zur Bestimmung der Computer- und Interneterfahrung zusammen (vgl. Abschn. 5.3.3). Alle Testpersonen verwenden im Rahmen ihrer Tätigkeit regelmäßig einen Computer. Davon geben knapp 90% an, in der Woche vor dem Experiment an fünf bis sieben Tagen mit einem Computer gearbeitet zu haben (M: 6,13; SD: 1,24). Die durchschnittliche Computernutzung beträgt 16,67 Stunden pro Woche (SD: 12,83), die durchschnittliche Internetnutzung 9,78 Stunden pro Woche (SD: 7,97). Wie man sieht, ist

Tab. 5.4.: Computer- und Interneterfahrung ($n = 89$).

Variable	Median	M	SD	Spanne
Computernutzung eine Woche vorher (Tage)	7	6,13	1,24	$3-7$
Computernutzung (Stunden pro Woche)	12,5	16,67	12,83	$2-60$
Internetnutzung (Stunden pro Woche)	8	9,78	7,97	$1-50$
Bekannte Suchmaschinen	2	2,56	1,34	$1-7$
Verwendete Suchmaschinen	1	1,66	0,94	$1-5$

der Unterschied zwischen Minimum und Maximum für diese beiden Kennzahlen recht groß, was zusammen mit den relativ hohen Werten für die Standardabweichung als Hinweis auf sehr unterschiedliche Erfahrungsniveaus in der Computer- und Internetnutzung verstanden werden kann. Schaut man sich jedoch an, wie häufig tatsächlich eine Stundenzahl größer als 20 Stunden pro Woche angegeben wird, zeigt sich, dass es sich hierbei im Fall der Internetnutzung eindeutig um Ausreißer handelt. So sind es lediglich fünf Teilnehmer (5,62%), die im Durchschnitt mehr als 21 Stunden in der Woche das Internet nutzen. Ohne Ausreißer beträgt die durchschnittliche Internetnutzung 8,22 Stunden pro Woche, die Standardabweichung verringert sich von 7,97 auf 4,44 Stunden, was für die Internetnutzungsfrequenz durchaus als realistisch angesehen werden kann. Hinsichtlich der Computernutzung ist zum einen anzumerken, dass es sich bei vier der Befragten, die eine durchschnittliche wöchentliche Computernutzung von über 20 Stunden pro Woche angeben, um Teilnehmer aus der Gruppe der Berufstätigen handelt, sodass eine höhere Stundenzahl in diesen Fällen vermutlich auf die Aufsummierung der beruflichen und der privaten Nutzungsdauer zurückzuführen ist. Einige Probanden weisen außerdem darauf hin, dass die Computernutzungsdauer während des Studiums stark davon abhängt, in welcher Studienphase man sich gerade befindet (Klausur, Referat, Hausarbeit etc.). Deshalb kommt die jeweils zutreffende Studienphase als weiterer Grund für die beobachteten Abweichungen in Frage. In Bezug auf die Bekanntheit und die Nutzung unterschiedlicher Suchdienste zeigt sich in dieser Stichprobe, dass die Befragten im Schnitt 2,56 unterschiedliche Suchmaschinen kennen (SD: 1,34), im Schnitt jedoch nur 1,66 regelmäßig verwenden (SD: 0,94). Die geringen Werte für die Standardabweichungen deuten darauf hin, dass in diesem Fall von einem vergleichbaren Wissensstand ausgegangen werden kann.

		System	
		gut	schlecht
Erwartung	niedrig	Gruppe 1 $n = 22$	Gruppe 2 $n = 22$
	hoch	Gruppe 3 $n = 22$	Gruppe 4 $n = 23$

Abb. 5.5.: Verteilung der Testteilnehmer auf die Untersuchungsgruppen ($n = 89$).

Zusammenfassend lässt sich die Stichprobe des ersten Experiments hinsichtlich der meisten soziodemographischen Kriterien als homogen bezeichnen. Hervorzuheben sind an dieser Stelle die Einschränkung der Stichprobe auf weibliche Testpersonen sowie der hohe Anteil studentischer Probanden. Die Verteilung der Probanden auf die vier Versuchsgruppen kann Abbildung 5.5 entnommen werden. In den nächsten beiden Abschnitten werden die zentralen Ergebnisse zur Suchleistung und zur Benutzerzufriedenheit berichtet. Dabei sind die folgenden Ausführungen an Werner (2010) angelehnt. Auf eine ausführlichere Darstellung der Ergebnisse wird verzichtet und stattdessen auf Lamm (2008) verwiesen.

5.4.2. Auswertung der Benutzerleistung

Zur Auswertung der fünf in Abschnitt 5.3.2 beschriebenen Leistungsmaße werden zweifaktorielle Varianzanalysen verwendet (vgl. Abschn. 4.3.2.1). Die Tabellen 5.5 und 5.6 zeigen die entsprechenden Ergebnisse auf Basis der erhobenen Untersuchungsdaten. Dabei wird die Benutzerleistung der Testpersonen über alle drei Aufgaben gemittelt. Tabelle 5.5 enthält die Ergebnisse in Bezug auf die Haupt- und Interaktionseffekte, in Tabelle 5.6 sind die entsprechenden Gruppenmittelwerte dargestellt. Nur bei der Pre-Click-Precision (PCP) ist die Normalverteilungsbedingung erfüllt, sodass die Nullhypothese für die übrigen Maße erst ab einer Irrtumswahrscheinlichkeit $p < 0{,}04$ verworfen wird (Zöfel, 2003, S. 217). Darüber hinaus ist die Varianzhomogenität bezüglich der Benutzer-Precision (BP) kritisch zu bewerten, weshalb die Irrtumswahrscheinlichkeit in diesem Fall auf $p < 0{,}01$ gesenkt wird (ebd., S. 217).

Tab. 5.5.: Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung. Signifikante Effekte sind fett hervorgehoben.

	Erwartungshaltung			**Systemgüte**			**Interaktion**		
Maß	**df**[a]	**F**[b]	**p**[c]	**df**	**F**	**p**	**df**	**F**	**p**
RRD	1	2,868	0,094	1	0,465	0,497	1	2,731	0,102
BR	1	2,562	0,113	1	0,519	0,473	1	2,316	0,132
T1	1	0,041	0,840	1	0,101	0,751	1	0,289	0,592
BP	1	0,486	0,488	1	13,045	**0,001**	1	4,823	**0,031**
PCP	1	0,939	0,335	1	4,424	**0,038**	1	1,013	0,317

[a] Freiheitsgrade [b] F-Wert [c] Signifikanz

Sowohl für die BP($F(1, 85) = 13{,}05$, $p < 0{,}005$) als auch für die PCP($F(1, 85) = 4{,}42$, $p < 0{,}05$) lässt sich ein signifikanter Effekt der Systemleistung nachweisen. In beiden Fällen erreichen Benutzer des besseren Systems höhere Benutzerleistungswerte. Für die übrigen Leistungsmaße lassen sich keine signifikanten Unterschiede zwischen den Versuchsgruppen feststellen. In Bezug auf die recallorientierten Maße scheinen Benutzer also, wie in früheren Studien berichtet (vgl. Abschn. 3.2.1), in der Lage zu sein, Systemunterschiede zu kompensieren. Insbesondere ist diese Beobachtung konsistent mit den Ergebnissen von Al-Maskari et al. (2008b) bzw. Al-Maskari et al. (2008a), da sich der relative Systemunterschied in Bezug auf die AvP mit 30% vs. 35% auf einem ähnlichen Niveau bewegt. Die erste Forschungshypothese (*Die Leistung der Benutzer wird durch die Systemgüte gemäß den Annahmen des systemorientierten Ansatzes positiv beeinflusst*) kann demnach nur z.T. als bestätigt angesehen werden.

Die Vermutung, dass Erwartungen eine zentrale Rolle für die Qualitätswahrnehmung von Suchergebnissen spielen, kann hingegen nicht bestätigt werden: Ein signifikanter Einfluss der Erwartungshaltung auf die Benutzerleistung ist nicht nachweisbar. Einerseits ist es möglich, dass

Tab. 5.6.: Gruppenmittelwerte der Benutzerleistungsmaße. Dargestellt sind die Mittelwerte in Abhängigkeit von System und Erwartung sowie der vier Vergleichsgruppen.

	Systemgüte		**Erwartungshaltung**		**Interaktion**			
	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$	$\mathbf{I_{G,H}}$	$\mathbf{I_{G,N}}$	$\mathbf{I_{S,H}}$	$\mathbf{I_{S,N}}$
RRD	9,55	8,75	8,16	10,14	7,59	11,5	8,73	8,77
BR	0,17	0,19	0,16	0,2	0,14	0,21	0,19	0,19
T1	444,38	440,23	443,62	440,99	442,2	446,56	445,04	435,42
BP	0,93	0,86	0,9	0,89	0,92	0,95	0,89	0,83
PCP	0,68	0,62	0,64	0,67	0,66	0,71	0,62	0,62

Erwartungen tatsächlich keinen Einfluss auf die Qualitätswahrnehmung ausüben. Dies würde jedoch früheren Studien widersprechen (vgl. Abschn. 2.1.1.3). Andererseits ist es aber auch möglich, dass die gedankliche Voreinstellung der Testpersonen durch die Testinstruktion nicht stark genug beeinflusst wird, um sich tatsächlich auf die Relevanzurteile der Versuchspersonen auszuwirken. Diese Vermutung wird im Anschluss an den Benutzertest durch einzelne Probanden bestätigt. Sie geben im informellen Gespräch zu, diesen Teil der Instruktion nur überflogen zu haben.

Die Haupteffekte bei BP und PCP sollen im Folgenden noch etwas eingehender betrachtet werden. Im Fall der PCP fällt auf, dass weder bei der Anzahl der richtig als relevant bewerteten Dokumente (RRD) noch bei der Anzahl der geöffneten Dokumente ein signifikanter Unterschied zwischen dem besseren und dem schlechteren System besteht. Erst das Verhältnis dieser beiden Werte ergibt sich ein signifikantes Ergebnis, da sich die Trends zu mehr geöffneten Dokumenten sowie zu weniger gefundenen relevanten Dokumenten im Falle des schlechteren Systems gegenseitig verstärken. Beim besseren System hingegen tritt der umgekehrte Effekt auf.

Das wohl interessanteste Ergebnis im Zusammenhang mit der Benutzerleistung ist der signifikante Unterschied in der BP. Um den Ursprung dieses Effekts zu ergründen, betrachtet man am besten die Definition der BP als Quotienten von RRD und der Anzahl der als relevant bewerteten Dokumente (RBD). Zerlegt man RBD in die Summe aus RRD und der Anzahl der fälschlicherweise als relevant markierten Dokumente (FRD), erhält man:

$$\mathrm{BP} = \frac{\mathrm{RRD}}{\mathrm{RBD}} = \frac{\mathrm{RRD}}{\mathrm{RRD} + \mathrm{FRD}} = \frac{1}{1 + \frac{\mathrm{FRD}}{\mathrm{RRD}}} \tag{5.1}$$

Die BP hängt somit ausschließlich von dem Quotienten $\frac{\mathrm{FRD}}{\mathrm{RRD}}$, also dem Verhältnis falsch als relevant und richtig als relevant bewerteter Dokumente, ab. Somit lässt sich aus dem Unterschied in der BP auf eine restriktivere bzw. weniger strenge Bewertung der Relevanz der Dokumente durch die Benutzer des besseren bzw. schlechteren Systems schließen. In Abbildung 5.6 sind diese unterschiedlichen Bewertungsstrategien noch einmal graphisch dargestellt. Es zeigt sich, dass Probanden aus der Versuchsgruppe mit dem schlechteren System eher dazu neigen, Dokumente fälschlich als relevant zu bewerten, als Testpersonen, die mit dem besseren System arbeiten. Der umgekehrte Effekt ist bei den fälschlich als irrelevant bewerteten Dokumenten zu beobachten.

Zusammenfassend lassen sich drei zentrale Ergebnisse festhalten. Erstens ist aus methodischer Perspektive zu vermuten, dass die experimentelle Manipulation der Erwartung nicht stark genug ausfällt, um die Relevanzurteile der Versuchspersonen zu beeinflussen. Zweitens lassen die nicht signifikanten Befunde der recallorientierten Benutzerleistungsmaße darauf schließen, dass Qua-

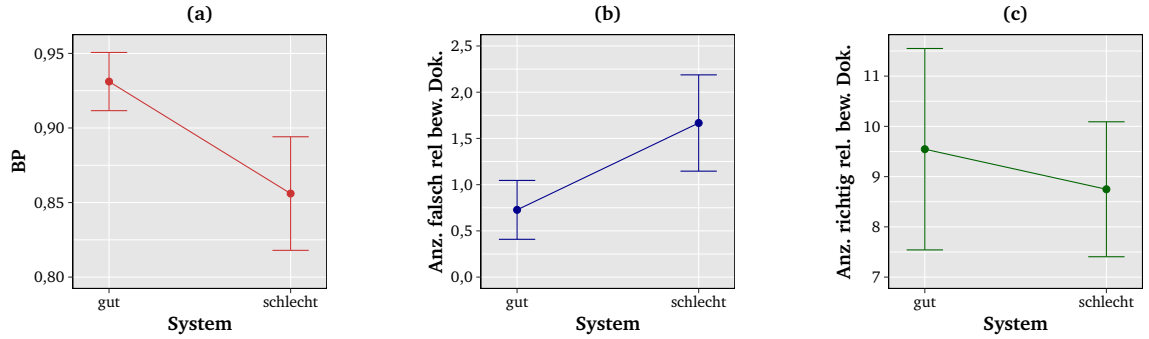


Abb. 5.6.: Systembedingte Anpassung der Relevanzwahrnehmung im ersten Experiment. Bild (a): Die Benutzerprecision ist für Nutzer des schlechteren Systems vermindert. Bild (b) und (c): Gegenläufiges Verhalten für die Anzahl der richtig bzw. fälschlicherweise als relevant bewerteten Dokumente. Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte.

litätsunterschiede beim Ranking durch die Benutzer kompensiert werden können. Dies deutet darauf hin, dass Benutzer in der Lage sind, trotz geringerer Systemleistung ihr Informationsbedürfnis zu befriedigen. In diesem Zusammenhang gewinnt die dritte Beobachtung zur Anpassung der Bewertungsstrategie besondere Bedeutung. Inwieweit eine höhere Suchmaschinenqualität tatsächlich zu strengeren Bewertungsmaßstäben bei der Relevanzbewertung führt, ist daher eine zentrale Fragestellung der Folgeuntersuchungen.

5.4.3. Auswertung der Benutzerzufriedenheit

Auch zur Auswertung der 15 Zufriedenheitsitems des ersten Experiments (vgl. Anh. A.3) werden zweifaktorielle Varianzanalysen durchgeführt. Zwar liegt in den meisten Fällen Varianzhomogenität vor, die Normalverteilungsvoraussetzung ist jedoch in keinem der Fälle erfüllt, sodass generell ein korrigiertes Signifikanzniveau von $p < 0{,}4$ zugrunde gelegt wird (Zöfel, 2003, S. 217). Die Varianzhomogenität ist in vier Fällen nicht gegeben (Item 2, Item 9, Item 11 und Item 14), weshalb die Nullhypothese für diese Items erst ab einer Irrtumswahrscheinlichkeit von $p < 0{,}01$ verworfen wird (ebd., S. 217).

Signifikante Gruppenunterschiede ergeben sich lediglich für zwei der 15 untersuchten Zufriedenheitsitems. Deshalb wird an dieser Stelle von einer tabellarischen Darstellung der gesamten Varianzanalyseergebnisse abgesehen und auf Lamm (2008, S. 90) verwiesen. Für die folgenden beiden Frageitems, die sich auf die Zufriedenheit mit der Precision der Suchergebnisse beziehen, kann ein signifikanter Haupteffekt der Systemleistung auf die Benutzerzufriedenheit nachgewiesen werden: Item 9 (*Die Artikel hätten besser gefiltert werden können.*) und Item 10 (*Die meisten Artikel waren für die dazugehörigen Suchanfragen relevant.*). Dabei ergibt die Varianzanalyse für Item 9 die Werte $F(1, 85) = 7{,}48$, $p < 0{,}01$ und für Item 10 die Werte $F(1, 85) = 5{,}22$, $p < 0{,}04$. Ein Vergleich der Gruppenmittelwerte zeigt in beiden Fällen, dass Probanden, die das bessere System verwenden, zufriedener sind als Benutzer des schlechteren Systems. Im Einzelnen ergeben sich für das bessere System die Werte $3{,}48$ für Item 9 und $5{,}05$ für Item 10. Für das schlechtere System liegen die Mittelwerte für Item 9 bei $2{,}53$ und für Item 10 bei $4{,}42$, wobei im Vergleich zu Lamm (ebd.) die Skala so gewählt wird, dass höhere Werte einer höheren Zufriedenheit entsprechen. Somit sind Benutzer also tatsächlich in der Lage, Unterschiede in der Systemleistung wahrzunehmen. Ein signifikanter Einfluss der Benutzererwartungen hingegen lässt sich, wie auch schon

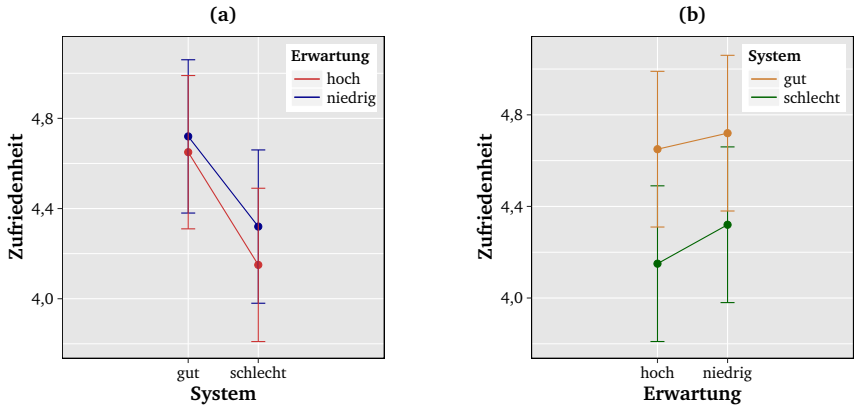


Abb. 5.7.: Interaktionsdiagramme zwischen Systemgüte und Erwartungshaltung für die Benutzerzufriedenheit. Bild (a): Systemeinfluss in Abhängigkeit der Erwartungshaltung. Bild (b): Erwartungseinfluss in Abhängigkeit der Systemgüte. Tendenziell lassen sich die Vorhersagen des C/D-Paradigmas erkennen, da sowohl die niedrige Erwartungshaltung als auch die bessere Systemleistung zu einer höheren Benutzerzufriedenheit führen. Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte

für die Benutzerleistungsmaße, für keines der 15 Frageitems nachweisen. Dies kann als weiterer Hinweis darauf gewertet werden, dass die Erwartungsmanipulation im ersten Experiment nicht stark ausfällt (vgl. Abschn. 5.4.2).

In einem zweiten Analyseschritt wird mit Hilfe der Reliabilitätsanalyse nach Cronbach (vgl. Abschn. 4.3.1) überprüft, ob sich diejenigen Zufriedenheitsitems, die sich ausschließlich auf die Retrievalqualität beziehen (Items 8 bis 14), zu einer gemeinsamen Skala zusammenfassen lassen. Die Reliabilitätsanalyse zeigt, dass ein Ausschluss von Item 12 (*Die Ergebnislisten waren zu umfangreich.*) eine Verbesserung des Cronbachs-Alpha-Werts auf $\alpha = 0{,}69$ nach sich ziehen würde. Die resultierende Zufriedenheitsskala beruht somit auf folgenden Frageitems: *Ich bin mit der Qualität der Suchergebnisse zufrieden.* (Item 8), *Die Artikel hätten besser gefiltert werden können.* (Item 9), *Die meisten Artikel waren für die dazugehörigen Suchanfragen relevant.* (Item 10), *Die Präsentation der Ergebnisse war übersichtlich.* (Item 11), *Die Reihenfolge der Suchergebnisse spiegelte die Relevanz der Artikel wider.* (Item 13) und *Die von mir aufgerufenen Artikel waren für die Recherche hilfreich.* (Item 14). Durch eine erneute Varianzanalyse auf der Basis dieser gemeinsamen Skala lässt sich der im Rahmen der Einzelanalysen gefundene Haupteffekt der Systemleistung bestätigen. Dabei ergeben sich die Werte $F(1, 85) = 6{,}94$, $p = 0{,}01$, wobei sowohl Normalverteilung als auch Varianzhomogenität gegeben sind. Die Erwartungshaltung wird auch im Fall der Gesamtskala nicht signifikant ($F(1, 85) = 0{,}47$, $p = 0{,}5$).

Abbildung 5.7 stellt die mittlere Zufriedenheit der vier Versuchsgruppen graphisch dar. Zumindest in der Tendenz lassen sich die Vorhersagen des C/D-Paradigmas erkennen (vgl. Abschn. 3.3.1.1), wenn auch der Einfluss der Erwartungshaltung hier nicht signifikant ist. Nichtsdestotrotz sieht man, dass Testpersonen mit einer niedrigen Erwartungshaltung im Mittel zufriedener sind als Testpersonen mit einer hohen Erwartungshaltung. Des Weiteren sind Benutzer des besseren Systems generell zufriedener als Benutzer des schlechteren Systems. Auf Grund der fehlenden Signifikanz in Bezug auf die Erwartungshaltung, kann die zweite Forschungshypothese (*Die Zufriedenheit der Benutzer wird durch ihre Erwartungshaltung und die Systemgüte gemäß den*

Annahmen des C/D-Paradigmas beeinflusst) jedoch nicht bestätigt werden.

Zusammenfassend lassen die nachgewiesenen Effekte der Systemleistung darauf schließen, dass Benutzer prinzipiell in der Lage sind, Qualitätsunterschiede in Bezug auf das Ranking wahrzunehmen. Die Vermutung, dass die Erwartungsmanipulation im ersten Experiment noch nicht stark genug ausgeprägt ist, wird durch die Auswertung der Benutzerzufriedenheit weiter bestätigt. Die Frage, inwieweit sich die Vorhersagen des C/D-Paradigmas tatsächlich auf den Prozess der Informationssuche übertragen lassen, kann daher an dieser Stelle nicht abschließend beantwortet werden und muss somit in den Folgestudien weiter untersucht werden.

5.4.4. Überprüfung der Gütekriterien des Experiments

Nachdem in den vorangegangenen beiden Abschnitten die wichtigsten Ergebnisse des ersten Experiments vorgestellt und in Bezug zu den forschungsleitenden Fragen gesetzt worden sind, werden in diesem Abschnitt die Gütekriterien des Experiments untersucht. In Bezug auf Faktoren, die das Urteil der Testpersonen über die Qualität der Suchmaschine möglicherweise zusätzlich beeinflussen, wird in diesem Zusammenhang zwischen untersuchungsbedingten und personenbezogenen Störfaktoren unterschieden. Erstere stehen für Faktoren, die auf das untersuchungsmethodische Vorgehen zurückzuführen sind. Letztere sind Faktoren, die primär mit den individuellen Merkmalen der Probanden in Zusammenhang stehen.

5.4.4.1. Untersuchungsbedingte Störfaktoren

Als mögliche untersuchungsbedingte Störfaktoren werden der Schwierigkeitsgrad der Aufgaben und ihre Bearbeitungsreihenfolge identifiziert. Darüber hinaus wird untersucht, ob eine Verkürzung der regulären Bearbeitungsdauer sowie die Tatsache, dass einige Probanden nur Dokumente auf der ersten Suchergebnisseite bewerten, eine Veränderung der Untersuchungsergebnisse nach sich ziehen.

Die Analyse möglicher Aufgaben- und Reihenfolgeeffekte ergibt, dass beide Effekte in den Daten vorhanden sind (Lamm, 2008, S. 82 ff.). In Bezug auf die Schwierigkeit der Aufgaben zeigt die Analyse, dass bis auf die BP alle Benutzerleistungsmaße einem signifikanten Aufgabeneffekt unterliegen. Im Vergleich scheint es etwas leichter zu sein das zweite Thema (Atomtransporte in Deutschland) zu bearbeiten als das erste (Erneuerbare Energien). Da auch in realen Anwendungskontexten nicht alle Aufgaben gleich schwer sind, kann dies jedoch als positive Voraussetzung für die Generalisierbarkeit der Untersuchungsergebnisse betrachtet werden. Signifikante Reihenfolgeeffekte zeigen sich nur bei dem ersten und dritten Thema (Kinderarbeit in Asien). Hier bestätigt sich die zuvor gemachte Beobachtung, dass das zweite Thema im Vergleich einfacher zu bearbeiten ist als die anderen beiden Aufgaben. Insgesamt betrachtet unterstützen diese Ergebnisse das untersuchungsmethodische Vorgehen. Insbesondere zeigt sich, dass die Randomisierung der Aufgabenreihenfolge aus methodischer Perspektive sinnvoll ist. Gleichzeitig scheint auch die Betrachtung der gemittelten Benutzerleistungswerte über alle drei Aufgaben gegenüber der Betrachtung der Einzelwerte der methodisch beste Weg zu sein, um ein möglichst genaues und unabhängiges Ergebnis zu erzielen.

Über alle Aufgaben hinweg wird in 40 der 267 Datensätze von der Möglichkeit Gebrauch gemacht, die Aufgabenbearbeitung vor Ablauf der regulären Bearbeitungszeit von zehn Minuten zu beenden, in 19 Datensätzen werden nur Dokumente auf der ersten Suchergebnisseite bewer-

tet. Die Überprüfung eines möglichen Einflusses dieser Faktoren ergibt, dass in beiden Fällen signifikante Unterschiede bestehen (ebd., S. 82 ff.). Diese deuten bezüglich einer Verkürzung der Bearbeitungszeit darauf hin, dass in diesen Fällen eine andere Suchstrategie verfolgt wird. Probanden, die kürzer suchen, scheinen generell etwas schneller zu arbeiten, da sich für zwei der drei Aufgaben ein signifikanter Einfluss auf die Zeit ergibt, die benötigt wird, um das erste relevante Dokument zu finden (T1). Weiterhin scheinen diese Probanden dazu zu tendieren, in ihrer Bewertung etwas weniger restriktiv vorzugehen, was sich an einer signifikant niedrigeren BP im Fall des ersten Themas zeigt. Probanden, die ausschließlich die erste Suchergebnisseite zur Beurteilung der Retrievalqualität heranziehen, erreichen erwartungsgemäß niedrigere Werte bei den recallorientierten Leistungsmaßen, da in diesen Fällen weniger relevante Dokumente zur Verfügung stehen. In Anbetracht der Tatsache, dass das Vorhandensein beider Effekte den Realitätsgrad der Untersuchung erhöht, sowie aufgrund der im Verhältnis geringen Fallzahlen wird in Bezug auf die in diesem Kapitel berichteten Untersuchungsergebnisse jedoch davon ausgegangen, dass diese Effekte vernachlässigbar sind.

5.4.4.2. Personenbezogene Störfaktoren

Als mögliche personenbezogene Störfaktoren werden die Muttersprache, das Alter und die Sucherfahrung der Probanden überprüft. Die in diesem Zusammenhang durchgeführten Analysen ergeben, dass weder die Muttersprache noch die Sucherfahrung der Probanden einen signifikanten Einfluss auf die abhängigen Variablen haben (ebd., S. 82 ff. u. S. 104 ff.), was dafür spricht, dass es sich bei der untersuchten Stichprobe tatsächlich um eine relativ homogene Gruppe handelt (vgl. Abschn. 5.4.1). Eine durchgeführte Kovarianzanalyse unter Einbeziehung des Alters als Kovariate ergibt zwar einen signifikanten Alterseinfluss bei den recallorientierten Leistungsmaßen, ohne jedoch einen Haupteffekt für Erwartung oder Systemleistung sichtbar werden zu lassen.

Zusammenfassend ist es im Zuge der Überprüfung untersuchungsbedingter Störfaktoren gelungen nachzuweisen, dass die verwendeten Suchaufgaben wie auch die Möglichkeit selbst zu entscheiden, wann die Aufgabe erfüllt ist, zu einem höheren Realitätsgrad des Untersuchungsdesigns führen. In diesem Zusammenhang wird deutlich, dass das gewählte untersuchungsmethodische Vorgehen zur Validität des Experiments beiträgt. Darüber hinaus scheinen als Störfaktoren vermutete personenbezogene Größen in diesem Experiment keinen nennenswerten Einfluss auf die in den Abschnitten 5.4.2 und 5.4.3 berichteten Ergebnisse zu besitzen.

5.5. Fazit: Experiment 1

Ausgangspunkt der Überlegungen zum ersten Experiment ist die Erkenntnis, dass Benutzer und ihre Interaktion mit dem System eine hohe Bedeutung für die Evaluierung von IR-Systemen besitzen, ihre wissenschaftliche Erforschung für den Bereich der IR-Evaluierung bisher allerdings vernachlässigt wurde. Deshalb ist es ein Ziel des ersten Experiments, die Übertragbarkeit des systemorientierten Ansatzes auf die Mensch-Maschine-Interaktion zu untersuchen. Dabei zeigt sich im Stand der Forschung, dass die Wahrnehmung und die Bewertung eines Produkts in der Kundenzufriedenheitsforschung eng mit den Erwartungen der Kunden verbunden ist (vgl. Abschn. 3.3.1). Angesichts der Vergleichbarkeit von Kauf- und Suchprozessen stellte sich daher zudem die Frage, ob dieser Einfluss auch für die Suchzufriedenheit gilt. Im Rahmen dises Experiments wird

deshalb ein benutzerorientiertes Untersuchungsdesign entwickelt, das die gleichzeitige Überprüfung dieser beiden Forschungsfragen gestattet.

Die Ergebnisse des ersten Experiments zeigen zunächst, dass das gewählte Untersuchungsdesign seine Grenzen hat, wenn es um die Manipulation der Erwartung geht. Die Frage, ob Erwartungen im Suchprozess tatsächlich keine Rolle spielen, oder ob die Erwartungsmanipulation nicht stark genug ist, kann an dieser Stelle nicht abschließend beantwortet werden. Für die Folgestudien sollte deshalb eine neue Manipulationsstrategie entwickelt werden. Dabei sollten folgende Punkte beachtet werden: Anstelle einer schriftlichen Instruktion der Testpersonen kann eine audiovisuelle Form der Darbietung treten. Im Gegensatz zur schriftlichen Form wird die Aufmerksamkeit der Testpersonen durch ein Einführungsvideo stärker fokussiert. Um darüber hinaus die gedankliche Voreinstellung der Versuchspersonen stärker zu beeinflussen, könnte es hilfreich sein, den Leistungsvergleich des C/D-Paradigmas aufzugreifen und für die Testpersonen sichtbar in das Untersuchungsdesign zu integrieren. Die Vermutung liegt nahe, dass die Einnahme der jeweiligen Erwartungshaltung auf diese Weise erleichtert wird. Zusätzlich sollte eine Kontrollfrage sicherstellen, dass die Manipulation wahrgenommen wird. Auch ohne signifikanten Einfluss der Erwartungshaltung sind die Vorhersagen des C/D-Paradigmas jedoch tendenziell in den Ergebnissen beobachtbar. Folgestudien müssen zeigen, ob sich dieser Trend erhärtet.

Die weiteren Ergebnisse zeigen jedoch, dass das Untersuchungsdesign im Übrigen geeignet ist, um die eingangs formulierten Forschungshypothesen (vgl. Abschn. 5.2) zu untersuchen. So kann in Übereinstimmung mit Turpin und Hersh (2001) gezeigt werden, dass Benutzer im Stande sind, Qualitätsunterschiede in Bezug auf das Ranking zu kompensieren und somit ihr Informationsbedürfnis auch bei geringerer Systemqualität erfüllen zu können. Demnach wäre der Mehrwert, den die Verwendung eines bestimmten IR-Systems seinem Anwender verschafft, weniger in der Effektivität des Systems, sondern vielmehr in der visuellen Aufbereitung der Suchergebnisse zu suchen. Die Beobachtung, dass Benutzer ihre Bewertungsstrategie der gegebenen Systemqualität anzupassen scheinen, spricht jedoch dafür, dass dem Ranking auch im benutzerorientierten Ansatz eine Schlüsselrolle zukommt. Gleichzeitig machen diese Ergebnisse deutlich, dass die Relevanzurteile von Benutzern kontextabhängig sind und deshalb im Hinblick auf die Evaluierung von IR-Systemen nicht isoliert betrachtet werden sollten. Ein Ziel der Folgestudien ist es deshalb, die Replizierbarkeit dieser Ergebnisse zu überprüfen.

Hinsichtlich der Zufriedenheit der Testpersonen mit der Retrievalqualität zeigt sich, dass Benutzer zwar die Fähigkeit haben, Qualitätsunterschiede zwischen IR-Systemen zu kompensieren, diese Unterschiede jedoch gleichwohl wahrzunehmen in der Lage sind. Diesbezüglich untermauert das erste Experiment die Ergebnisse von Al-Maskari et al. (2007, S. 773), die ebenfalls einen Zusammenhang zwischen Precision und Benutzerzufriedenheit nachweisen. Mit Blick auf die Folgestudien ist zu überlegen, den Fragenkatalog in Bezug auf Items zur Erfassung der Zufriedenheit mit der Retrievalqualität auszudehnen, da der Einfluss der Systemleistung bei diesen Frageitems besonders ausgeprägt ist. Des Weiteren scheint es empfehlenswert, zusätzlich ein bereits etabliertes Fragebogeninstrument hinzuzuziehen, um auf diese Weise die Vergleichbarkeit mit anderen Studien zu erhöhen.

6. Experiment 2: Steuerbarkeit der Qualitätswahrnehmung

Aufbauend auf den Ergebnissen der im vorangegangenen Kapitel vorgestellten Nutzerstudie wird im Jahr 2009 ein zweites Experiment durchgeführt, bei welchem primär die Eignung der Benutzererwartung als zentrale Steuerungsgröße der Benutzerzufriedenheit untersucht wird. Im Zentrum stehen dabei erneut die in Abschnitt 1.2 beschriebenen Forschungsfragen FF1 und FF2. Zum einen soll die Übertragbarkeit des C/D-Paradigmas auf den Kontext der Informationssuche weiter erforscht werden. Zum anderen wird auf die Frage eingegangen, wie sich ein Unterschied in der Systemleistung auf das Suchverhalten von Benutzern auswirkt. Um den Vergleich der beiden Experimente zu erleichtern, entspricht der Aufbau des Kapitels weitgehend der aus Kapitel 5 bekannten Gliederung.

6.1. Untersuchungsziel

Da die Erwartungsmanipulation des ersten Experiments keinen statistisch signifikanten Einfluss auf die Suchleistung und die Zufriedenheit der Versuchspersonen zeigt, besteht ein wesentliches Ziel des zweiten Experiments darin, die Manipulation der Erwartungshaltung zu verstärken. Die Idee, die hinter dem neuen experimentellen Untersuchungsdesign steht, beruht darauf, dass ein Vergleich zweier unterschiedlicher Systeme dabei helfen kann die Erwartungsmanipulation zu verstärken. Zum einen kann so der Anlass des Benutzertests realistischer kommuniziert werden. Zum anderen erscheint es hilfreich, den Leistungsvergleich des C/D-Paradigmas aufzugreifen und für die Testpersonen sichtbar in das Untersuchungsdesign zu integrieren, da den Testpersonen die Einnahme der jeweiligen Erwartungshaltung auf diese Weise erleichtert wird. Eine Kontrollfrage am Ende des Benutzertests soll außerdem Hinweise auf den Erfolg der Erwartungsmanipulation liefern, um das eigentliche Anliegen der Studie nicht deutlich werden zu lassen. Eine weitere Beobachtung im Kontext des ersten Experiments betrifft die Tatsache, dass einzelne Testpersonen die Testinstruktionen und somit auch die Erwartungsmanipulation, nicht aufmerksam genug lesen. Um dieser Problematik im zweiten Experiment entgegenzuwirken, erfolgt die Testinstruktion dieses Mal audiovisuell mit Hilfe eines Einführungsvideos. Sowohl die gleichzeitige Ansprache des Hör- und Sehsinns als auch die emotionalere Form der Testinstruktion soll die Aufmerksamkeit der Testpersonen während der Erwartungsmanipulation erhöhen. Ein direkter Vergleich der ersten beiden Experimente findet sich außerdem in Lamm et al. (2010b).

6.2. Forschungsleitende Hypothesen

Nachdem im Zuge des ersten Experiments gezeigt werden kann, dass objektive Systemunterschiede nicht zwangsläufig zu messbaren Verbesserungen in Bezug auf die Benutzerleistung führen

(vgl. Abschn. 5.4.2), konzentrieren sich die ersten beiden Forschungshypothesen dieser Nutzerstudie auf die im ersten Experiment beobachteten Verhaltenseffekte. Insbesondere deutet die Abwesenheit signifikanter Haupteffekte darauf hin, dass Benutzer für recallorientierte Benutzerleistungsmaße in der Lage zu sein scheinen, Systemleistungsunterschiede zu kompensieren. Mit der ersten Forschungshypothese soll die Replizierbarkeit dieser Beobachtung überprüft werden.

H1: Bei recallorientierten Leistungsmaßen können Benutzer Unterschiede in der Systemgüte kompensieren.

Die zweite Beobachtung bezieht sich auf die im ersten Experiment eingesetzten precisionorientierten Benutzerleistungsmaße. Die Auswertung der Benutzerprecision lässt vermuten, dass Benutzer ihre persönliche Relevanzdefinition an der vorgefundenen Retrievalqualität ausrichten. Es zeigt sich, dass die Relevanzurteile der Testpersonen umso restriktiver ausfallen, je höher die Qualität der dargebotenen Ergebnislisten ist (vgl. Abschn. 5.4.2). Dieser Verhaltenseffekt soll mit der zweiten Forschungshypothese weiter untersucht werden.

H2: Bei precisionorientierten Leistungsmaßen passen Benutzer ihre Relevanzdefinition der Systemgüte an.

Die Tatsache, dass das C/D-Paradigma im ersten Experiment zwar in der Tendenz sichtbar, jedoch nicht statistisch signifikant ist, führt zu der Annahme, dass die Erwartungsmanipulation im ersten Experiment möglicherweise nicht stark genug ausfällt. Aus diesem Grund bleibt die erste Forschungshypothese des ersten Experiments vorerst bestehen. Es wird weiterhin von der Eignung des C/D-Paradigmas als Erklärungsmodell der Suchmaschinenzufriedenheit ausgegangen, sodass die dritte Forschungshypothese wie folgt lautet:

H3: Die Zufriedenheit der Benutzer wird durch ihre Erwartungshaltung und die Systemgüte gemäß den Annahmen des C/D-Paradigmas beeinflusst.

6.3. Methode

Zur Überprüfung der Hypothesen wird ein experimentelles Untersuchungsdesign entworfen, das den Vergleich beider Systemvarianten ermöglicht. Dabei wird den Testpersonen mitgeteilt, dass sie zwei unterschiedliche IR-Systeme im Vergleich bewerten sollen und jeweils eine Suchaufgabe mit einem der beiden Systeme bearbeiten sollen. Als Konsequenz hieraus ergibt sich der in Abbildung 6.1 dargestellte Versuchsplan mit acht verschiedenen Untersuchungsgruppen. Wie auch im ersten Experiment erfolgt die Zuordnung der Probanden zu den einzelnen Gruppen nach dem Zufallsprinzip. Entsprechend dem Versuchsplan erhalten bspw. die Teilnehmer in Gruppe 4 die Information, dass sie zuerst das bessere und anschließend das schlechtere System testen werden, tatsächlich handelt es sich bei den präsentierten Ergebnislisten jedoch in beiden Fällen um die schlechtere Systemvariante. Im Gegensatz zum ersten Experiment verwenden die Teilnehmer dieses Mal also zwei vermeintlich unterschiedliche Suchsysteme. Die Messung der wahrgenommenen Retrievalqualität erfolgt dabei wie im ersten Experiment anhand verschiedener in Abschnitt 6.3.2 beschriebener Zufriedenheits- und Benutzerleistungsmaße, die jeweils pro Aufgabe erfasst werden. Die angestrebte Stichprobengröße liegt bei insgesamt 160 Probanden bzw. 20 Teilnehmern pro Untersuchungsgruppe.

			System							
			A_1 gut	A_2 gut	A_1 gut	A_2 schlecht	A_1 schlecht	A_2 gut	A_1 schlecht	A_2 schlecht
Erwartung	A_1	hoch	Gruppe 1		Gruppe 2		Gruppe 3		Gruppe 4	
	A_2	niedrig								
	A_1	niedrig	Gruppe 5		Gruppe 6		Gruppe 7		Gruppe 8	
	A_2	hoch								

Abb. 6.1.: Versuchsplan des zweiten Experiments. In dem gewählten Between-Subjects-Design führt die Nutzung der beiden Testsysteme zu den dargestellten acht Untersuchungsgruppen. A_1 und A_2 bezeichnen jeweils die Erwartungshaltung und Systemgüte bei der ersten bzw. zweiten Suchaufgabe. Für jede Aufgabe wird sowohl die Benutzerzufriedenheit als auch die Benutzerleistung gemessen.

6.3.1. Manipulation der unabhängigen Variablen

Die Systemqualität und die Benutzererwartungen stellen auch im zweiten Experiment die unabhängigen Variablen dar. Um die Vergleichbarkeit der Experimente untereinander zu ermöglichen, werden die für das erste Experiment gewählten Systemleistungsunterschiede beibehalten. Anders als beim ersten Experiment haben die Teilnehmer diesmal jedoch die Möglichkeit, ihre Suchanfragen frei zu wählen und ggf. auch umzuformulieren (vgl. Abschn. 6.3.3). Ähnlich wie bei Turpin und Scholer (2006, S. 14 f.) bekommen die Testpersonen im Falle einer erneuten Suche eine alternative Ergebnisliste mit gleicher Systemleistung präsentiert. Diese Ergebnislisten werden im Vorfeld des Experiments mit Hilfe des bereits in Abschnitt 5.3.1 eingeführten Algorithmus von Turpin und Scholer (ebd.) erzeugt, sodass, insgesamt 100 Listen pro Systemvariante zur Verfügung stehen. Die verwendeten Rankinglisten finden sich in Anhang B.1. Die randomisierte Zuteilung der relevanten und irrelevanten Dokumente erfolgt zur Laufzeit während der Untersuchung. Auch dieses Mal wird bei der Erstellung der Ergebnislisten zusätzlich die Precision innerhalb der ersten zehn Dokumente kontrolliert, damit die Manipulation der Systemleistung auch bei einer für die Internetsuche typischen Einschränkung auf die vorderen Rankingpositionen zu Tage tritt. Dabei wird jedoch keine strikte Anzahl an irrelevanten Dokumenten mehr vorgegeben, sondern darauf geachtet, dass ein überwiegender Teil der schlechteren Ergebnislisten mindestens vier irrelevante Dokumente innerhalb der ersten zehn Dokumente enthält, die $P@10$ also maximal $0{,}6$ beträgt. Im Vergleich dazu werden die Ergebnislisten für das bessere System so gewählt, dass der Großteil der Listen maximal zwei irrelevante Dokumente innerhalb der ersten zehn Dokumente enthält, also eine $P@10$ von mindestens $0{,}8$ aufweist. Die genaue Verteilung der irrelevanten Dokumente in den verwendeten Ergebnislisten kann Abbildung 6.2 entnommen werden.

Um die Konsistenz der unterschiedlichen Ergebnislisten weiter zu analysieren, wird im Folgenden noch die Verteilung der Effektivitätsmaße Discounted Cumulative Gain (DCG), Normalized

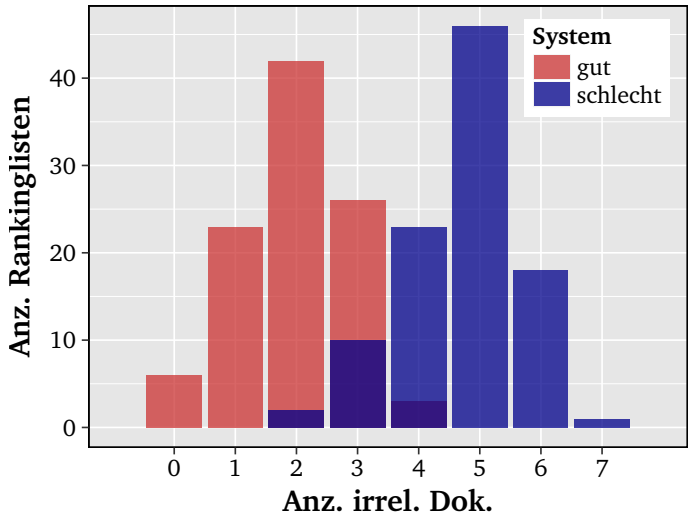


Abb. 6.2.: Anzahl irrelevanter Dokumente innerhalb der ersten 10 Treffer.

Discounted Cumulative Gain (nDCG) und Binary Preference (BPref) überprüft (vgl. Abschn. 4.2.2.2). Anhand von Abbildung 6.3 zeigt sich für alle Maße sowohl eine gute Trennung zwischen den beiden Systemvarianten als auch eine gute Übereinstimmung mit geringer Streuung innerhalb der beiden Systemgüten.

Um den Einfluss der Manipulation der Erwartungshaltung auf die Bewertung der Retrievalqualität im zweiten Experiment zu erhöhen, wird eine andere Manipulationsstrategie gewählt, bei der alle Testpersonen gebeten werden, zwei Systeme im Vergleich zu bewerten. Das Ziel dieses Vorgehens liegt darin, den Probanden die Einnahme der jeweiligen Erwartungshaltung zu erleichtern. Eine weitere Änderung betrifft die Darbietung der Testinstruktion. Anstelle der Papierform bekommen die Probanden im zweiten Experiment ein Einführungsvideo gezeigt. Durch diese emotionalere Form der Testinstruktion soll die Aufmerksamkeit der Testpersonen während der Erwartungsmanipulation erhöht werden. Außerdem wird angenommen, dass die gleichzeitige Ansprache des Hör- und Sehsinns eine unterstützende Funktion bei der Aufnahme der Erwartungsmanipulation erfüllt. Im Rahmen des Einführungsvideos wird den Probanden zunächst mitgeteilt, dass im Rahmen eines Projekts mit dem Namen Infofokus zwei Suchmaschinen entwickelt wurden, die in ersten Systemtests unterschiedlich gut abgeschnitten haben. Als Erwartungsmanipulation erfahren die Testpersonen weiterhin, dass die blaue Suchmaschine im Durchschnitt mehr relevante und gleichzeitig weniger irrelevante Dokumente gefunden hat als die grüne. Die farbliche Kennzeichnung wird dabei aus einer Studie von Kelly et al. (2007) übernommen. Sie dient dazu, den Testpersonen die Unterscheidung zwischen den beiden Suchmaschinen zu erleichtern und gleichzeitig stets präsent zu halten, dass es sich um einen Vergleich von zwei unterschiedlichen Systemen handelt. Als Ziel des Experiments wird vorgegeben, die Qualität der beiden Suchmaschinen aus Benutzersicht zu bewerten. Der genaue Wortlaut der Testinstruktion ist Anhang B.2 zu entnehmen.

Durch eine Kontrollfrage am Ende des Tests soll diesmal überprüft werden, ob die Erwartungsmanipulation erfolgreich ist. Nur Probanden, die diese Kontrollfrage richtig beantworten, werden ohne Einschränkungen in die Auswertung einbezogen. Die Kontrollfrage für die Erwartungsmanipulation lautet: *Wissen Sie noch, welche der beiden Suchmaschinen in den statistischen*

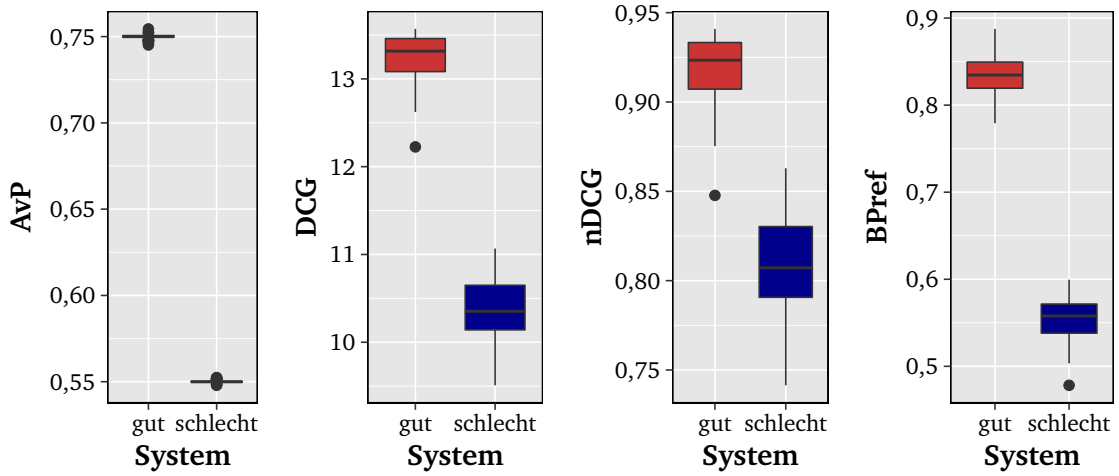


Abb. 6.3.: Qualität der Ergebnislisten bezüglich weiterer Effektivitätsmaße.

Tests besser abgeschnitten hat? Neben dieser eindeutig definierten Kontrollfrage wird auch die aus der Nutzung resultierende Einstellung zu den beiden Suchsystemen protokolliert. Um diese zu erfassen enthält der Zufriedenheitsfragebogen im Anschluss an die zweite Aufgabe 12 zusätzliche Frageitems. Zunächst werden die Untersuchungsteilnehmer gebeten, für eine weitere Suchaufgabe die von ihnen präferierte Suchmaschine zu wählen und ihre Suchleistung anhand der voraussichtlich in zehn Minuten auffindbaren relevanten Dokumente einzuschätzen. Die nächsten vier Items sind der Studie von Szajna und Scamell (1993, S. 516) zu Informationssystemen entnommen (vgl. Abschn. 4.2.1.2) und an den vorliegenden Kontext angepasst. Die Items sind als Fragen formuliert (*Wie wahrscheinlich ist es, dass ... ?*) und sollen anhand einer 7-stufigen, bipolaren Skala von *sehr unwahrscheinlich* bis *sehr wahrscheinlich* beantwortet werden. Inhaltlich behandeln diese Items zum einen universelle Zusammenhänge zwischen Handlung und Ergebnis der Suche (*Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?*). Zum anderen umfassen sie jedoch auch Items mit einem hohen Selbstbezug (*Wie wahrscheinlich ist es, dass Sie von der Leistung, die sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?*), die Aspekte der in Abschnitt 2.2.2 beschriebenen Selbstwirksamkeitserwartungen mit einschließen. Der genaue Wortlaut aller Items ist in Anhang B.4 aufgeführt.

6.3.2. Operationalisierung der abhängigen Variablen

Nach der Beschreibung des Stimulusmaterials im vorangegangenen Abschnitt werden nun die Erhebungsinstrumente zur Erfassung der abhängigen Variablen vorgestellt. Analog zum ersten Experiment stellen die Verhaltens- und Zufriedenheitsreaktionen der Testteilnehmer auch in dieser Untersuchung die abhängigen Variablen dar. Über die bereits dargestellten inhaltlichen Ziele hinaus wird im Rahmen des zweiten Experiments ein größeres Gewicht auf die Validität der Ergebnisse zur Nutzerzufriedenheit gelegt. Aus diesem Grund das von Doll und Torkzadeh (1988) entwickelte EUCS-Instrument als standardisierter Fragebogen zur Erfassung der Benutzerzufriedenheit eingesetzt (vgl. Abschn. 4.2.2.3). Es umfasst 14 bipolare Items mit je fünf Antwortkategorien (*fast nie - manchmal - in der Hälfte der Fälle - meistens - fast immer*), die zu folgenden fünf faktorenanalytisch begründeten Standardskalen zusammengefasst werden: Inhalt, Genauigkeit,

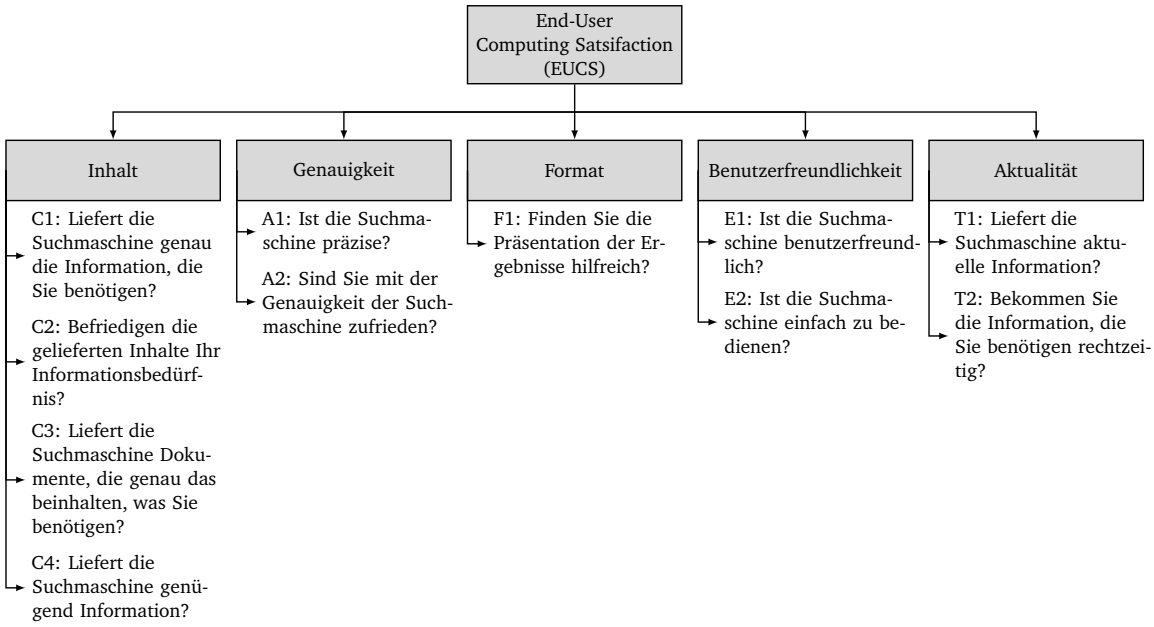


Abb. 6.4.: Zufriedenheitsfaktoren und Frageitems des EUCS-Instruments (nach Doll und Torkzadeh, 1988, S. 268). Das im Originalinstrument enthaltene Item F2: *Is the information clear?*, wird mangels einer präzisen Übersetzung, die eng am englischen Original bleibt, nicht berücksichtigt, um die Validität des Instruments nicht zu gefährden.

Darstellung, Benutzerfreundlichkeit und Aktualität. Die einzelnen Frageitems werden zunächst mit Hilfe eines Muttersprachlers aus dem Englischen übersetzt und anschließend im Rahmen des Pretests auf seine Eignung hin überprüft (vgl. Abschn. 6.3.7). Dabei wird darauf geachtet, dass die Übersetzung der Frageitems sich soweit wie möglich an den Original-Items orientiert, um die Validität des Fragebogens nicht zu gefährden. In diesem Zusammenhang wird entschieden, das Darstellungsitem F02 (*Is the information clear?*) im Rahmen dieser Untersuchung nicht zu erheben, da eine wörtliche Übersetzung in diesem Fall unklar erscheint. Abbildung 6.4 zeigt das dem EUCS-Instrument zugrunde liegende Modell zur Messung der Benutzerzufriedenheit. Neben den elf in Abbildung 6.4 dargestellten Items enthält das EUCS-Instrument außerdem zwei weitere Items die als Außenkriterium zur Validierung des Fragebogens eingesetzt werden können (vgl. Abschn. 6.4.4.1).

Um den Schwerpunkt bei der Erhebung der Zufriedenheit noch deutlicher auf die Qualität der Suchergebnisse sowie die zu bearbeitende Suchaufgabe zu legen, werden ergänzend selbst entwickelte Zusatzitems verwendet, sodass die Zufriedenheit der Testpersonen im zweiten Experiment insgesamt anhand von 26 Items gemessen wird. Mit dem Ziel die Vergleichbarkeit zur Studie aus dem Jahr 2008 zu erhöhen, werden die beiden signifikanten Items des ersten Experiments (Item 09 u. 10) übernommen und an das Untersuchungsdesign angepasst. Wie im ersten Experiment soll außerdem weiterhin die Möglichkeit bestehen, die Zufriedenheit der Testpersonen auf indirekte Art und Weise zu erfassen. Zu diesem Zweck werden die Untersuchungsteilnehmer im letzten Item des Fragebogens gebeten anzugeben, inwiefern sie es sich vorstellen können, die getestete Suchmaschine als Standardsuchmaschine in ihrem Browser einzustellen. Auch die im vorherigen Abschnitt beschriebenen Items zur Messung der aus der Erfahrung resultierenden Erwartung können als indirekte Items zur Zufriedenheitsmessung angesehen werden. Bei der

Konstruktion der weiteren Items kann u.a. auf Vorarbeiten von Kelly et al. (2008b) zurückgegriffen werden. Im Gegensatz zu den 13 Items des EUCS-Instruments, werden die 13 Zusatzitems wie im ersten Experiment anhand einer 7-stufigen Bewertungsskala beurteilt, um auch hier eine Vergleichbarkeit zwischen den ersten beiden Experimenten herzustellen. Um diesmal jedoch eine konsistente Beantwortung zu ermöglichen, wird die Richtung der Skala umgedreht, sodass wie im Fall des EUCS-Instruments eine höhere Zustimmung durch einen höheren Skalenwert repräsentiert wird. Eine Liste aller verwendeten Items findet sich in Anhang B.3.

Die Suchleistung der Testpersonen wird anhand der in Abschnitt 4.2.2.2 eingeführten Benutzerleistungsmaße erfasst. Im Unterschied zu Experiment 1 ist jedoch die Möglichkeit einer iterativen Suche zu berücksichtigen: Die Testpersonen können ihre Suchanfragen frei wählen und ggf. umformulieren. Dies führt dazu, dass die Teilnehmer zwar Ergebnislisten gleicher Systemgüte aber potentiell unterschiedliche Dokumente präsentiert bekommen (vgl. Abschn. 6.3.1). Dies bedeutet insbesondere, dass dasselbe Dokument im Kontext unterschiedlicher Suchanfragen auftauchen und entsprechend von den Testpersonen bewertet werden kann. Auf der einen Seite verringert dies zwar potentiell die Homogenität zwischen den einzelnen Versuchsteilnehmern, auf der anderen Seite erhöht dieses Vorgehen jedoch auch den Realismus des Versuchsaufbaus. Dabei werden bei der praktischen Berechnung der Leistungsindikatoren, wie in Abschnitt 4.2.2.2 dargestellt, die bewerteten Dokumente der gesamten Session zugrunde gelegt. Die Anzahl der richtig als relevant bewerteten Dokumente (RRD) ergibt sich somit bspw. als Vereinigung über die richtig als relevant bewerteten Dokumente über alle gestellten Suchanfragen hinweg, wobei für jedes Dokument ausschließlich die letzte Bewertung berücksichtigt wird. Die Gesamtanzahl der relevanten Dokumente bezüglich welcher bspw. der Benutzerrecall berechnet wird, ist dementsprechend durch die Gesamtzahl der relevanten Dokumente im Testkorpus oder alternativ durch die Zahl der in der gesamten Session zurückgelieferten relevanten Dokumente gegeben.

Abgesehen von den bereits in Experiment 1 berücksichtigten Standardmaßen, werden im zweiten Experiment eine Vielzahl spezifischerer Leistungsmaße zur Beurteilung der Benutzerleistung herangezogen. In Abbildung 6.5 sind schematisch einige dieser, über die Standardmengen hinausgehende, Differenzierungsmöglichkeiten zur Leistungsbeurteilung graphisch zusammengefasst. Dabei wird zunächst allgemein auf die von einer Testperson während der Suche aufgerufenen und bewerteten Dokumente zurückgegriffen, was zu einer ersten Aufteilung in relevant und irrelevant bewertete Dokumente führt. Da darüber hinaus für jedes Dokument Jurorenurteile bezüglich einer binären Relevanzskala vorliegen (vgl. Abschn. 6.3.4), lassen sich weiterhin für das gegebene Thema relevante und irrelevante Dokumente unterscheiden. Ausgehend von diesen einfachen Teilmengen eröffnet sich nun ein breites Spektrum potentieller Leistungsmaße. So lässt sich bspw. untersuchen, wie sich die durchschnittliche Bewertung relevanter aufgerufener Dokumente in Abhängigkeit von Erwartungshaltung und Systemgüte ändert. Ebenso lässt sich analysieren, wie viele der vom Nutzer als relevant bewerteten Dokumente von den Juroren als relevant oder irrelevant eingestuft werden. Darüber hinaus ist es auch möglich, spezifische Phasen des Suchprozesses genauer zu analysieren und die zu untersuchende Untermenge bspw. auf die erste bzw. letzte durchgeführte Suche oder die ersten zehn Rankingplätze einzuschränken. Dies erlaubt bspw. konkret zu untersuchen, wie viele der ersten zehn pro Suche zurückgelieferten Dokumente in Übereinstimmung mit den Juroren als irrelevant bewertet werden. Ausgehend

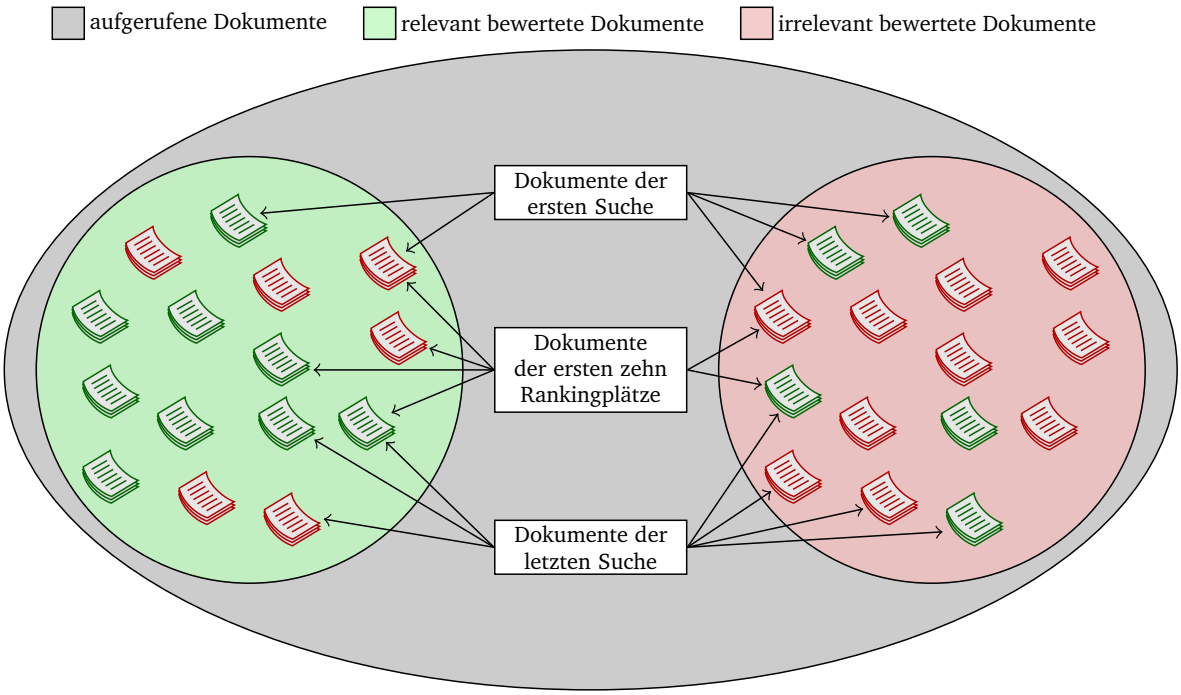


Abb. 6.5.: Schematische Darstellung ausgewählter Untermengen der aufgerufenen Dokumente. Innerhalb der Menge der vom Nutzer als relevant bzw. irrelevant bewerteten Dokumente sind bspw. die im Rahmen der ersten Suche aufgerufenen Dokumente zusätzlich markiert. Von den Juroren als relevant annotierte Dokumente sind als grüne, irrelevant bewertete Dokumente hingegen als rote Dokumentenpiktogramme dargestellt.

von diesen Dokumentenmengen lassen sich weitere Verhältnis- und Zeitmaße ableiten, um die Benutzerleistung zu quantifizieren.

Konkret können die verwendeten Leistungsindikatoren in fünf Gruppen untergliedert werden: Die erste Gruppe umfasst die eingangs beschriebenen Dokumentenmengen. Dazu zählen z.B. die Anzahl der aufgerufenen relevanten Dokumente, die Anzahl der als relevant bewerteten Dokumente, die Anzahl der richtig als relevant bewerteten Dokumente, aber auch die Anzahl der im Rahmen der letzten durchgeführten Suche als relevant bewerteten Dokumente oder die Anzahl der als relevant bewerteten Dokumente, die auf den ersten zehn Rankingplätzen angezeigt werden. Insgesamt werden im zweiten Experiment 19 verschiedene Dokumentenmengen berücksichtigt. Die zweite Gruppe beinhaltet die durchschnittliche Relevanzbewertung innerhalb einzelner dieser Dokumentenmengen. Konkret handelt es sich bspw. um die durchschnittliche Bewertung der aufgerufenen relevanten oder irrelevanten Dokumente während der gesamten Suchsitzung oder in einer speziellen Phase (z.B. erste Suche). Die im zweiten Experiment erhobenen binären Relevanzurteile werden dabei als 1 (relevant bewertet) bzw. 0 (irrelevant bewertet) gezählt, sodass die sechs in dieser Gruppe berücksichtigten Leistungsmaße alle einen Wertebereich zwischen 0 und 1 aufweisen. Die dritte Gruppe enthält die durchschnittlichen Betrachtungszeiten in Sekunden für 10 der in der ersten Gruppe definierten Dokumentenmengen. Darüber hinaus wird die durchschnittliche Betrachtungsdauer eines beliebigen Dokuments erhoben. Da Betrachtungszeiten nur selten normalverteilt sind (Kelly u. Belkin, 2004; Yin et al., 2013; Rummel, 2014), dies jedoch eine der Voraussetzungen varianzanalytischer Methoden darstellt, werden die hier betrachteten Mittelwerte einmal von logarithmierten und einmal von nicht

logarithmierten Einzelzeiten gebildet, sodass jedes Leistungsmaß dieser Gruppe in zweifacher Ausführung in die Auswertung eingeht. In der vierten Gruppe werden verschiedene Verhältnisse bzw. Quotienten von Dokumentenmengen betrachtet. Neben den in Abschnitt 5.3.2 bereits erläuterten Standardmaßen BR, BP und PCP enthält diese Gruppe 33 weitere Leistungsmaße, die im Zusammenhang mit der Untersuchung des Einflusses von Systemleistung und Erwartungshaltung eine Rolle spielen könnten. Diese umfassen insbesondere die in Abschnitt 4.2.2.2 eingeführten Imprecisionmaße, die darauf abzielen zu quantifizieren, wie stark die Probanden mit ihrem Relevanzurteil von den Bewertungen der Juroren abweichen. Es zeigt sich, dass gerade diese Maße geeignet sind, die Reaktion der Testteilnehmer auf Systemunterschiede in Bezug auf ihre Relevanzwahrnehmung zu detektieren. Insgesamt erlaubt diese Gruppe von Leistungsindikatoren somit einen präziseren Blick auf einzelne Aspekte der klassischen Recall- und Precisionmaße. In die fünfte und letzten Gruppe schließlich fallen fünf weitere Leistungsmaße, die nicht in die übrigen Kategorien einzuordnen sind. Darunter fallen die bereits im ersten Experiment untersuchte Zeit bis zum ersten richtig relevant bewerteten Dokument, die Dauer der gesamten Suche, die Anzahl der durchgeführten Suchen sowie die erste und die letzte betrachtete Rankingposition. Eine detaillierte Liste aller 86 im zweiten Experiment zum Einsatz kommenden Benutzerleistungsvariablen befindet sich in Anhang B.9.

Zusammenfassend kann gesagt werden, dass das Ziel einer möglichst breiten und lückenlosen Erfassung des Verhaltens und Zufriedenheitsempfindens der Probanden mit den hier vorgestellten Erhebungsinstrumenten erreicht werden kann. Mit dem übersetzten EUCS-Instrument zur Erfassung der Benutzerzufriedenheit steht ein standardisierter Fragebogen zur quantitativen Bewertung der subjektiv wahrgenommenen Qualität von Suchmaschinen zur Verfügung. Dieses wird darüber hinaus noch um weitere Frageitems erweitert, die z.T. ebenfalls auf bereits erprobten Messinstrumenten basieren. Auch hinsichtlich der auf den Sucherfolg der Probanden bezogenen Leistungsvariablen kann eine breite Abdeckung des Konstruktinhalts gewährleistet werden. Die Einbeziehung der üblichen Standardmaße zur Erfassung der Benutzerleistung sichert auch hier die Validität der Daten und stellt gleichzeitig die Vergleichbarkeit der Ergebnisse sicher.

6.3.3. Umgang mit Störvariablen

Ähnlich wie im ersten Experiment wird ein Einfluss auf die abhängigen Variablen durch zusätzliche Störvariablen angenommen. Im Folgenden wird beschrieben, welche Maßnahmen im Rahmen des zweiten Experiments getroffen werden, um ihren Einfluss zu kontrollieren. Sowohl aus Gründen der Vergleichbarkeit als auch um Alterseffekte auszuschließen, werden die Teilnehmer des zweiten Experiments aus der gleichen Altersgruppe wie im ersten Experiment rekrutiert. Von einer Konstanthaltung des Geschlechts wird diesmal mit Blick auf den angestrebten größeren Stichprobenumfang abgesehen (vgl. Abschn. 6.3). Stattdessen werden die Untersuchungsteilnehmer geschlechtshomogen randomisiert. Zusätzlich wird ein möglicher Einfluss von Alter, Geschlecht und Muttersprache auf die Ergebnisse durch eine Einbeziehung als Kovariate überprüft.

Ein besonderes Anliegen des zweiten Experiments besteht darin, ein größeres Gewicht auf den Einfluss des Erfahrungswissens der Untersuchungsteilnehmer zu legen. Neben der Sucherfahrung der Probanden wird diesmal auch ihr inhaltliches Wissen in Bezug auf die zu bearbeitenden

PageRank

PageRank ist ein Verfahren zur Bewertung der Wichtigkeit von Webseiten, bei dem davon ausgegangen wird, dass...

a) eine Webseite umso relevanter ist, je mehr Einzelseiten sie hat. ☐

b) eine Webseite umso relevanter ist, je mehr externe Links auf den Seiten zu finden sind. ☐

c) eine Webseite umso relevanter ist, je mehr interne Links auf den Seiten zu finden sind. ☐

d) eine Webseite umso relevanter ist, je mehr andere Homepages auf sie verweisen. ☐

weiß nicht ☐

Abb. 6.6.: Frageitem zum theoretischen Suchmaschinenwissen.

Suchaufgaben abgefragt. Als Ausgangslage werden die Teilnehmer gebeten, eine Selbsteinschätzung ihrer Kompetenz zu den behandelten Themengebieten abzugeben. Die Selbsteinschätzung wird anhand der Kompetenzstufen Anfänger, Fortgeschrittener und Experte vorgenommen. Um neben der Befragung der Teilnehmer auch ein objektives Maß für das Erfahrungswissen der Testpersonen zu erhalten, werden zwei Wissenstests entwickelt, die im Folgenden kurz vorgestellt werden. Die Tests enthalten zwischen 8 und 11 Items, bei denen jeweils nur eine der vorgegebenen Antwortmöglichkeiten zutrifft. Angelehnt an das Inventar zur Computerbildung (INCOBI) von Richter et al. (2001b) sind die einzelnen Items als Quizfragen umgesetzt. Bei beiden Wissenstests ergibt sich der Skalenwert eines Probanden aus der Summe richtig beantworteter Items. Wie auch im Fall des Inventars zur Computerbildung sind die Items des Sucherfahrungsfragebogens so konzipiert, dass sowohl theoretisches als auch praktisches Wissen im Zusammenhang mit Suchmaschinen erfragt wird. Im Gegensatz zu Fragen wie *Verwenden Sie Boolesche Operatoren (AND, OR und NOT), um ihre Suchstrategie zu verbessern?*, die unabhängig vom Kenntnisstand beantwortet werden könnten, ermöglicht es die Wahl von Quizfragen in stärkerem Maße, das tatsächliche Wissen der Probanden zu erfassen. Für die Konstruktion der Items kann auf verschiedene im Internet verfügbare Quiz zurückgegriffen werden[1], die für die Wissenstests angepasst werden. Außerdem können einzelne Fragen aus einer Studie von Hölscher (2000) übernommen oder an dort verwendete Formulierungen angelehnt werden. Die Abbildungen 6.6 und 6.7 zeigen zwei Beispielitems. Das erste Beispiel (vgl. Abb. 6.6) ist eine typische Quizfrage, mit der das

[1] Verwendet werden Items aus folgenden Quellen (zuletzt geprüft am 05.07.2010):
Wissenstest: Erneuerbare Energien, `http://www.geo.de/GEO/interaktiv/wissenstests/52472.html`; Energiespar-Wissenstest: Ist Ihr Haus gewappnet für die Zukunft? `http://www.rb-hs.de/ware_gewerbe/energie_weiter/energiespar-wissenstest0.html`; Solar-Quiz: Was leistet Sonnenenergie? `http://www.focus.de/panorama/jahresrueckblick-2004/quiz/solar-quiz_aid_12201.html`, Pfadfinder für das weltweite Netz, `http://portal.gmx.net/de/themen/digitale-welt/quiz/4055724.html`; Quiz: Suchen und Suchmaschinen, `http://www.internet-abc.de/eltern/quiz-suchmaschinen.php`, Wissens-Test Informationsbeschaffung im Internet, `http://swisseduc.ch/informatik/internet/internet_recherche/test/`

theoretische Wissen der Testpersonen im Zusammenhang mit Suchmaschinen getestet wird. Das zweite Beispiel (vgl. Abb. 6.7) erhebt das praktische Wissen der Testpersonen auf diesem Gebiet. Die vollständigen Itemlisten der beiden Wissenstests finden sich im Anhang (vgl. Anh. B.5 und B.6). Die im ersten Experiment verwendeten Items zur Erfassung der Sucherfahrung werden aus Gründen der Vergleichbarkeit der beiden Untersuchungen in abgewandelter Form ebenfalls mit erhoben (vgl. Anh. B.7).

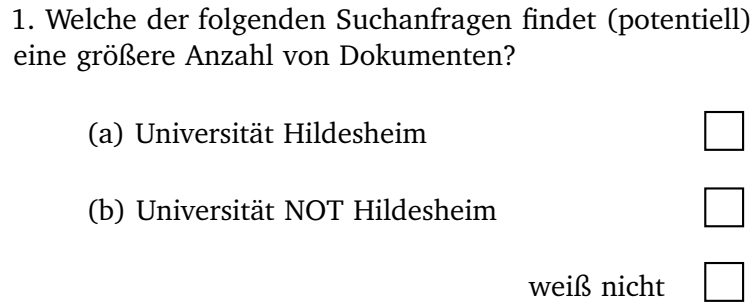


Abb. 6.7.: Frageitem zum praktischen Suchmaschinenwissen.

Weiterhin wird davon ausgegangen, dass Studierende IT-orientierter Studiengänge per se über ein größeres Erfahrungswissen im Umgang mit Suchmaschinen verfügen. Aus diesem Grund werden aus diesen Studiengängen nur Studienanfänger zur Teilnahme zugelassen. Da im zweiten Experiment mit jedem der beiden Systeme nur eine Suchaufgabe bearbeitet wird, ist hier eine Kontrolle möglicher Lern- und Reihenfolgeeffekte besonders wichtig. Aus diesem Grund wird ein möglicher Einfluss des Suchthemas durch Parallelisierung kontrolliert, indem jedes Thema in jeder Versuchsgruppe sowohl als erste als auch als zweite Aufgabe verwendet wird.

Als weitere personenbedingte Störvariable wird die Motivation der Testteilnehmer kontrolliert. Gemäß Heckhausen und Heckhausen (2010, S. 7) hängt die Motivation einer Person für eine bestimmte Aktivität „[. . .] von situativen Anreizen, persönlichen Präferenzen und deren Wechselwirkung ab.“ Im vorliegenden Experiment wird deshalb ein situativer Anreiz durch einen in Aussicht gestellten Gewinn für die beste Suchleistung geschaffen. Diesbezüglich können sich die Teilnehmer am Ende des Tests entscheiden, welche der beiden Aufgaben für den Wettbewerb gewertet werden soll (vgl. Abschn. 6.3.6). Darüber hinaus dient diese Auswahl als weiteres indirektes Zufriedenheitsurteil. Außerdem werden alle Versuchspersonen am Anfang und am Ende des Experiments gebeten, ihre momentane Befindlichkeit auf einer 5-stufigen Skala von -2 bis $+2$ einzuschätzen.

In Bezug auf die Simulation des Suchprozesses wird im zweiten Experiment, in Anlehnung an Turpin und Scholer (2006), auf die Vorgabe von Suchbegriffen verzichtet. Dies führt mit Blick auf die Relevanzbewertung zu einem gewissen Verlust an experimenteller Kontrolle, da die Relevanzurteile der Probanden nun nicht mehr ausschließlich von der Suchaufgabe sondern auch von der gewählten Suchanfrage abhängig sein können. Gleichzeitig jedoch erscheint dieser Kontrollverlust zugunsten eines realistischeren Sucherlebnisses und somit einer höheren externen Validität des Experiments in gewissem Maße unvermeidbar. Auch die Tatsache, dass sich die Ergebnislisten abgesehen von der jeweiligen Systemleistung für alle Testpersonen unterscheiden und darüber hinaus beliebig viele Suchanfragen eingegeben werden dürfen, sorgt dafür (vgl. Abschn. 6.3.1), dass sich die Variabilität im zweiten Experiment zugunsten einer realistischeren

Simulation des Suchprozesses erhöht. Bei der Auswertung muss daher zusätzlich die Plausibilität der eingegebenen Suchbegriffe überprüft werden.

Ein weiterer Unterschied zum ersten Experiment besteht in dem Verfahren der Datenerhebung. Während im ersten Experiment ausschließlich Einzeltests zur Anwendung kommen, ermöglicht der neue serverbasierte Ansatz (vgl. Abschn. 6.3.5) im zweiten Experiment die zeitgleiche Durchführung mehrerer Tests in einem Raum, was eine wesentliche Zeitersparnis darstellt. Zu beachten ist, dass durch die gleichzeitige Testdurchführung neue Störfaktoren wie z.B. wechselseitige Störungen der Testpersonen untereinander oder die Gefahr des gegenseitigen Abschreibens hinzukommen. Da die Testpersonen außerdem auch im zweiten Experiment bis zu einer maximalen Dauer von zehn Minuten selbst entscheiden können, wann sie die Bearbeitung der Aufgabe beenden, kann es zudem auch hier durch das Gruppenverfahren zu einer gegenseitigen Beeinflussung kommen. Der entscheidende Vorteil gegenüber Einzeltestverfahren besteht jedoch in der höheren Ökonomie von Gruppenverfahren. Bei der Platzierung in den PC-Arbeitsräumen wird deshalb darauf geachtet, dass die Testpersonen nach Möglichkeit nicht zu dicht sitzen. Außerdem werden alle Teilnehmer zu Beginn instruiert sich während des Test ruhig zu verhalten, um andere nicht zu stören.

Damit werden im zweiten Experiment insgesamt 12 Kovariaten identifiziert, die sich den Bereichen demographische Merkmale, Vorwissen, und Motivation zuordnen lassen. Dies erlaubt es, eine große Bandbreite an möglichen Störeinflüssen auf die abhängigen Variablen Benutzerzufriedenheit und Benutzerleistung zu kontrollieren, um wie in den Forschungsfragen formuliert den separaten Einfluss von Erwartungshaltung und Systemgüte herausarbeiten zu können. Eine Übersicht über alle im zweiten Experiment berücksichtigten Kovariaten kann Anhang B.10 entnommen werden.

6.3.4. Aufbau des Testkorpus

Aufgrund des gewählten Untersuchungsdesigns wird auch im zweiten Experiment ein Testkorpus benötigt, das es erlaubt, während des Experiments die Systemleistung zu manipulieren und im Nachgang die Suchleistung der Testpersonen bestimmen zu können. Aus Aktualitätsgründen sowie aus Gründen der praktischen Relevanz der Suchthemen wird auf die Nachnutzung einer standardisierten Dokumentensammlung verzichtet. Stattdessen wird im Rahmen des zweiten Experiments ein neues Testkorpus entwickelt, bei dem zu Gunsten einer hohen externen Validität ein besonderes Gewicht auf eine möglichst realitätsnahe Untersuchungssituation gelegt wird. Hierbei kann ähnlich wie in früheren Studien (vgl. Abschn. 4.1.3.3) auf Dokumente aus dem Internet zurückgegriffen werden. Neben der hohen Aktualität der Dokumente besteht ein weiterer Vorteil eines solchen Korpus in seiner Größe und dem unkomplizierten Zugang über Suchmaschinen. Ein Problem könnte in der Unbeständigkeit der Daten bestehen. Dies kann jedoch durch ein Speichern der Webseiten umgangen werden. Die im Zusammenhang mit der Erstellung des neuen Testkorpus bedeutsamen Arbeitsschritte werden in diesem Abschnitt erläutert.

Bei der Formulierung der Testaufgaben ist, wie im Fall des ersten Experiments, auf einen homogenen Schwierigkeitsgrad der Aufgaben zu achten sowie darauf, dass die Suchdomäne ein breites Publikum anspricht. Zusätzlich zu den in Abschnitt 5.3.4 genannten Kriterien, wird auch auf die Wiederverwendbarkeit der Testaufgaben geachtet. So wird bspw. versucht, tagesaktuelle Themen zu vermeiden. In Anlehnung an die erste Aufgabe des ersten Experiments wird der

Themenbereich *Erneuerbare Energien* als Oberthema gewählt. Konkret lauten die im zweiten Experiment verwendeten Suchthemen wie folgt:

1. Stellen Sie sich vor, Sie überlegen eine solarthermische Anlage zur Heizungsunterstützung anzuschaffen. Informieren Sie sich im Detail über mögliche Förderungen.

2. Stellen Sie sich vor, in der Nähe Ihres Wohnortes soll eine Windkraftanlage gebaut werden. Informieren Sie sich im Detail über die Geräuschentwicklung von Windkraftanlagen.

Zur Beurteilung der Relevanz wird wie im ersten Experiment eine binäre Unterteilung in relevante und irrelevante Dokumente gewählt. Diese Bewertung erfolgt durch zehn Juroren (einer davon ist die Verfasserin dieser Arbeit, im Folgenden als Juror 1 bezeichnet). Um Reihenfolgeeffekte zu vermeiden, wird den Juroren pro Suchthema eine randomisierte Liste von Dokumenten präsentiert. Jedes Dokument wird durch mindestens zwei unabhängige Juroren bewertet. Kann dabei keine Einigung erreicht werden, wird eine dritte Meinung herangezogen. Den Juroren ist keine bestimmte Bewertungsmethode vorgegeben. Stattdessen werden sie instruiert möglichst so zu entscheiden, wie sie es auch in einer realen Suchsituation tun würden. Tabelle 6.2 zeigt die Verteilung der relevanten und irrelevanten Dokumente innerhalb des erstellten Korpus. Pro Aufgabe umfasst die Kollektion 109 bzw. 111 Dokumente, wovon in beiden Fällen 45 Dokumente als irrelevant bewertet sind. Für beide Aufgaben werden im Schnitt für jedes Dokument circa $2{,}6$ unabhängige Relevanzurteile benötigt, um sie als relevant oder irrelevant zu klassifizieren. Dabei bewertet jeder Juror im Schnitt etwa 19 Dokumente pro Aufgabe.

Um die Konsistenz des Testkorpus einschätzen zu können, wird als nächstes die Interrater-Reliabilität zwischen den Juroren bestimmt (vgl. Abschn. 4.1.3.3). Abgesehen von Juror 1, der alle im Korpus enthaltenen Dokumente beurteilt, bewerten die übrigen Juroren maximal 40 Dokumente pro Aufgabe. Dies führt dazu, dass mit Ausnahme von Juror 1 keine nennenswerten Überschneidungen in Bezug auf die bewerteten Dokumente zwischen den einzelnen Juroren vorliegen, weshalb Cohens Kappa ausschließlich im Vergleich zu den Bewertungen von Juror 1 berechnet wird. Dabei kann Cohens Kappa die Werte zwischen -1 und 1 annehmen. Je näher der Wert an 1 liegt, desto größer ist die Übereinstimmung zwischen den Jurorenurteilen (vgl. Abschn. 4.1.3.3). Die resultierenden Kappa-Werte sind in Tabelle 6.1 dargestellt.

Zunächst fällt ein deutlicher Unterschied zwischen den beiden Suchthemen auf. Mit einem Durchschnittswert für Cohens Kappa von knapp $0{,}5$ scheint das erste Thema (Sonnenenergie) deutlich reliabler als das zweite Thema (Windenergie) zu sein ($\kappa = 0{,}2$). Für das erste Thema ergibt sich somit eine moderate Übereinstimmung des Urteils des ersten mit allen anderen Juroren (vgl. Abschn 4.1.3.3). Teilweise finden sich auch Übereinstimmungen von über $0{,}6$. Das im Vergleich schlechtere Abschneiden für das zweite Thema relativiert sich jedoch etwas wenn man neben Cohens Kappa noch weitere Kennzahlen berücksichtigt. So ist bspw. der Anteil der Dokumente bei denen eine dritte Bewertung notwendig ist mit $28\,\%$ (Thema 1) versus $35\,\%$ (Thema 2) von ähnlicher Größenordnung. Gleiches gilt für den Anteil der Dokumente bei denen das Urteil des ersten Jurors bestätigt wird, der mit $86\,\%$ beim zweiten Thema nur leicht unter der Rate von $92\,\%$ beim ersten Thema liegt. Eine mögliche Ursache für die geringen Kappa-Werte im Fall des zweiten Themas könnte in der Korrektur für zufällige Übereinstimmungen begründet liegen (vgl. Abschn. 4.1.3.3). So beträgt bspw. die unkorrigierte Übereinstimmung zwischen den

Tab. 6.1.: Interrater-Reliabilität zur Konsistenzprüfung des Testkorpus. Angegeben sind sowohl die erreichten Werte für Cohens Kappa als auch die unkorrigierten prozentualen Übereinstimmungswerte. Darüber hinaus schließt der korrigierte Mittelwert für das Windenergiethema die Juroren 2, 6, 7 und 9 aus.

	Sonne			Wind		
	Bew. Dok.	**Kappa**	**rel. Übereinst.**	**Bew. Dok.**	**Kappa**	**rel. Übereinst.**
Juror 2	18	0,26	0,72	19	0,00	0,89
Juror 3	37	0,52	0,76	39	0,53	0,77
Juror 4	35	0,62	0,83	34	0,38	0,71
Juror 5	20	0,44	0,75	15	0,35	0,67
Juror 6	17	0,49	0,76	18	0,03	0,56
Juror 7	9	0,61	0,89	12	0,03	0,50
Juror 8	19	0,58	0,79	20	0,47	0,75
Juror 9	8	0,29	0,67	19	-0,19	0,58
Mittelwerte	20	0,48	0,77	22	0,20	0,68
				korrigiert:	0,43	

Tab. 6.2.: Beschreibung des verwendeten Testkorpus. Für jedes Thema sind die Dokumentenanzahlen aufgeschlüsselt nach relevanten und irrelevanten Dokumenten angegeben. Des Weiteren ist die mittlere Anzahl notwendiger Jurorenurteile pro Dokument vermerkt, um dieses als relevant bzw. irrelevant zu klassifizieren. Da mindestens zwei übereinstimmende Jurorenurteile notwendig sind, beträgt die minimale Urteilsanzahl 2.

Thema	**Dok.**	**Anz. Dok.**	**⌀ Anz. Urteile**
Wind	irrelevant	45	2,20
	relevant	66	2,32
	gesamt	111	2,27
Sonne	irrelevant	45	2,20
	relevant	64	2,27
	gesamt	109	2,24
Gesamt	irrelevant	90	2,20
	relevant	130	2,29
	gesamt	220	2,25

ersten beiden Juroren bezüglich des zweiten Themas $89\,\%$. Da der erste Juror in diesem Fall allerdings fast alle von beiden Juroren bewerteten Dokumente als relevant gekennzeichnet hat, ist die Wahrscheinlichkeit einer zufälligen Übereinstimmung so hoch, dass sich insgesamt $\kappa = 0$ ergibt. In Tabelle 6.1 sind deshalb sowohl die ermittelten Kappa-Werte als auch die unkorrigierten prozentualen Übereinstimmungswerte angegeben. Weiterhin ist zu bemerken, dass die Juroren 6 und 7 fast ausschließlich Dokumente bewerten, bei denen schon zwei sich widersprechende Jurorenurteile vorliegen. Somit ist davon auszugehen, dass es sich in diesen Fällen um schwierig einzuschätzende Dokumente handelt, was das geringe Cohens Kappa von je $0{,}03$ erklären würde. Im Fall von Juror 9 für den Cohens Kappa sogar negativ wird, ist zu bemerken, dass seine von Juror 1 abweichenden Bewertungen nur in zwei Fällen durch einen dritten Juror bestätigt werden. Schließt man diese vier Juroren aus der Mittelwertbildung aus, ergibt sich ein mittleres Kappa von $0{,}43$, was sich in der gleichen Größenordnung wie beim ersten Thema bewegt und somit einer moderaten Übereinstimmung entspricht. Zusammenfassend kann also durchaus von einer konsistenten Bewertung der Relevanz der einzelnen Dokumente für beide Themen ausgegangen werden.

Um schließlich sicherzustellen, dass alle Testpersonen während des Experiments dasselbe Stimulusmaterial erhalten, müssen die im Korpus enthaltenen Dokumente archiviert werden.

Der Download der Internetseiten erfolgt mittels WinHTTrack[2], einer Software mit der komplette Webseiten heruntergeladen und gespeichert werden können. Nach Ausschluss einiger in Bezug auf das Speichern mittels WinHTTrack problematischer Dokumente verbleiben im finalen Korpus für das erste Thema 62 relevante und 45 irrelevante Dokumente und für das zweite Thema 63 relevante und 45 irrelevante Dokumente. Da pro Ergebnisliste nur 45 (schlechtes System) bzw. 54 (gutes System) relevante Dokumente benötigt werden, führt dies zu einer Variation der angezeigten Dokumente bei unterschiedlichen Suchanfragen, was noch einmal den Realitätsgrad der Testsituation erhöht.

6.3.5. Beschreibung des Testsystems

Um im Vergleich zum ersten Experiment den Realitätsgrad des Testsystems zu erhöhen, wird im Rahmen des zweiten Experiments ein browserbasiertes Testsystem entwickelt. Wie bei einer tatsächlichen Suchmaschine interagieren die Probanden dabei mit einer Webseite, die die Suchanfragen an einen Server sendet, der die Ergebnislisten ausliefert. Bei der Gestaltung der Webseite wird sich an der UX gängiger Suchmaschinen orientiert. Ein Screenshot des Testsystems mit angezeigter Suchergebnisliste ist in Abbildung 6.8 dargestellt. Wie in Abschnitt 6.3.1 bereits beschrieben, wird vom Testsystem für jede neue Suchanfrage eine neue Ergebnisliste angezeigt. Um die Illusion eines echten Suchgefühls zu erzeugen und eine wirklichkeitsgetreue Suchprozesssimulation zu ermöglichen, werden identische Suchanfragen derselben Person durch das Testsystem erkannt und mit derselben Ergebnisliste beantwortet. Auch die Trefferbeschreibungen mit Titel und Snippet entsprechen in ihrem Erscheinungsbild typischen Trefferdarstellungen. Sowohl Titel als auch Snippet werden im Rahmen der Korpuserstellung von Google übernommen. Die Benutzeroberfläche des Testsystems liegt dabei in zweifacher Ausführung vor, je nach Erwartungsmanipulation in blau (besseres System) oder in grün (schlechteres System). Im Gegensatz zu einem realen Suchsystem führen auch die Links in der Ergebnisliste zu auf dem Server gehosteten Seiten, die zuvor per WinHTTrack gespeichert werden. Ein Nachteil dieses Vorgehens liegt darin, dass diese URLs in der Statusleiste des Browsers angezeigt werden. Allerdings ist diese Anzeige so unscheinbar, dass nicht davon auszugehen ist, dass den Testpersonen diese Verlinkung zurück auf den Testserver auffällt. Abbildung 6.9 zeigt das Testsystem nach der Auswahl eines Links. In der linken Navigationsleiste gibt es nun die Möglichkeit, die Relevanz des angezeigten Dokuments zu bewerten.

Technisch ist das Testsystem als eine Sammlung von PHP-Skripten realisiert, die mit einer MySQL-Datenbank interagieren. In der Datenbank werden sowohl die Logdaten gespeichert als auch die Adressen der gehosteten Webseiten sowie ihre Relevanz in Bezug auf die beiden Suchthemen vorgehalten. Die Implementierung der Suchseite und ihre Interaktion mit dem Server basiert auf HTML und Javascript. Ein großer Vorteil dieses serverbasierten Ansatzes ist die Möglichkeit, mehrere Tests parallel durchführen und gleichzeitig die Logdaten zentral in einer Datenbank ablegen zu können. Zum Festlegen der Untersuchungsbedingungen wird für jede Versuchsgruppe eine andere URL auf dem Server aufgerufen (start1.php, start2.php, ...). Daher handelt es sich, wie im Fall des ersten Experiments, um einen Blindversuch, d.h. die Zuteilung zu den Versuchsgruppen ist zwar nicht der Testperson, wohl aber dem Testleiter bekannt.

[2] Bei WinHTTrack handelt es sich um die Windowsversion der GPL-Software HTTrack, die es erlaubt Offline-Kopien von Webseiten zu erzeugen (`https://www.httrack.com/`).

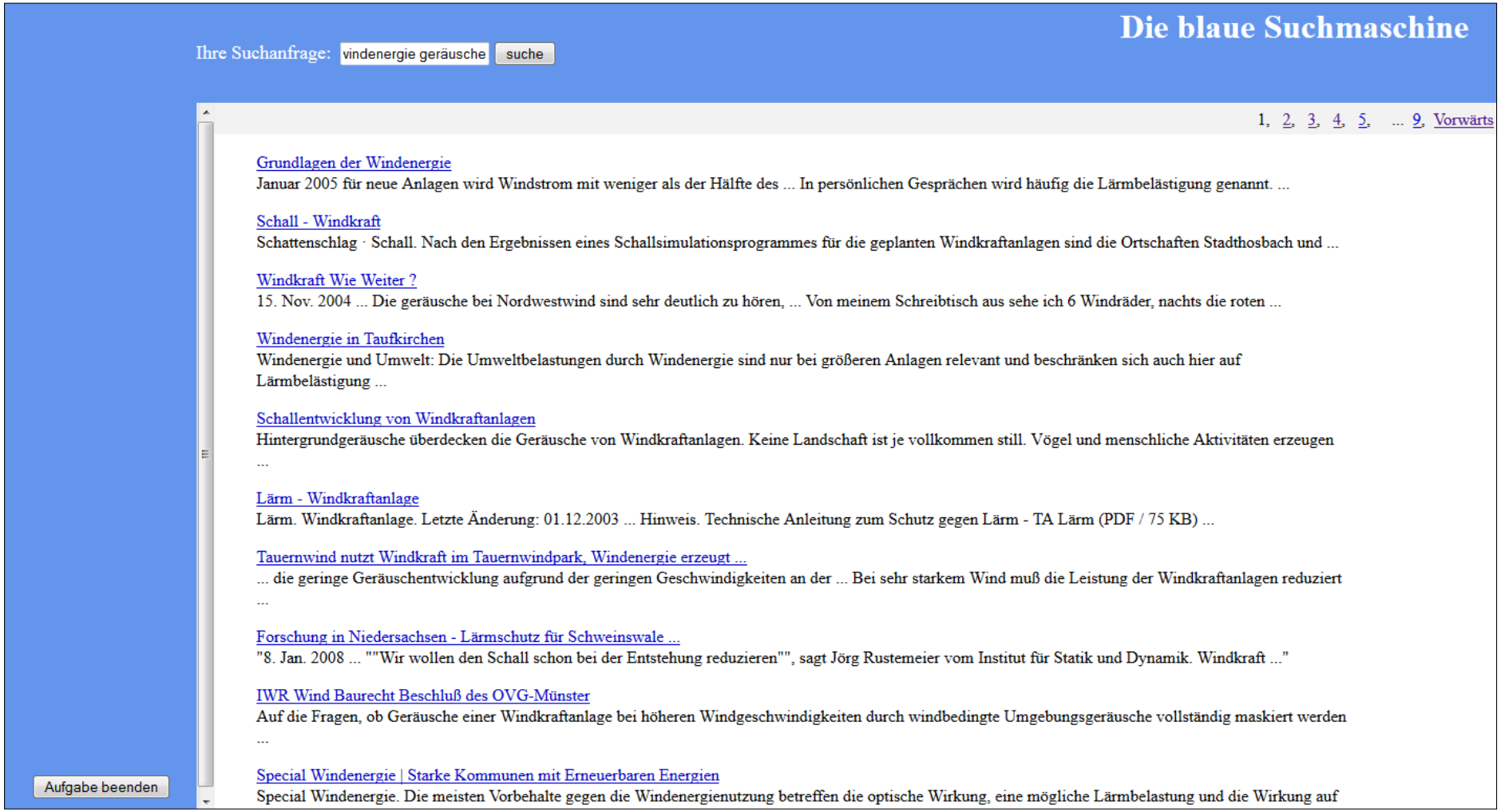


Abb. 6.8.: Testsystem des zweiten Experiments: Darstellung der Suchergebnisliste

Abb. 6.9.: Testsystem des zweiten Experiments: Relevanzbewertung

6.3.6. Ablauf

Das zweite Experiment ist wesentlich durch den Vergleich der beiden Systeme bestimmt. Der genaue Ablauf wird in Abbildung 6.10 beschrieben. Zur Schaffung einheitlicher Bedingungen für alle Probanden, findet das zweite Experiment in den PC-Arbeitsräumen der Universität Hildesheim statt. Zusammen mit dem in Abschnitt 6.3.5 beschriebenen serverbasierten Ansatz ermöglicht es dieser Aufbau, mehrere Tests parallel durchzuführen. Auch im zweiten Experiment ist der Ablauf für alle Teilnehmer identisch. Das Experiment beginnt mit einem Einführungsvideo, in dem neben der Erwartungsmanipulation eine Beschreibung des Ablaufs der Untersuchung sowie der Bedienung des Testsystems erfolgt. Auch im zweiten Experiment besteht jederzeit die Möglichkeit, offene Fragen im Einzelgespräch zu klären. Alle Versuchspersonen werden zu Beginn und am Ende des Tests gebeten, ihre momentane Befindlichkeit einzuschätzen (vgl. Abschn. 6.3.3). Anschließend erfolgt die Bearbeitung des ersten Wissenstests (vgl. Abschn. 6.3.3). Dieser dient dazu herauszufinden, über welches Vorwissen zum Thema erneuerbare Energien die Probanden zu Beginn des Experiments bereits verfügen. Gleichzeitig soll er den Testpersonen den Einstieg in das Thema erleichtern. Der letzte Teil des Tests beinhaltet außerdem einige allgemeine Fragen zur Person. Als nächstes folgt der praktische Teil. Um die für die Teilnahme erforderliche Zeit in vertretbaren Grenzen zu halten, sind im zweiten Experiment nur zwei Aufgaben zu bearbeiten, je eine pro Suchmaschine. Abhängig von der Versuchsgruppe erhalten die Teilnehmer dabei in den beiden Aufgaben Ergebnislisten gleicher oder unterschiedlicher Systemqualität (vgl. Abschn. 6.3). Sowohl die Abfolge der Systeme als auch die Reihenfolge der Aufgaben wird systematisch variiert (vgl. Abschn. 6.3.3). Um den Realitätsgrad des Versuchsaufbaus gegenüber dem ersten Experiment zu erhöhen, können Untersuchungsteilnehmer des zweiten Experiments ihre Suchbegriffe frei wählen und umformulieren (vgl. Abschn. 6.3.3). Jede neue Suchanfrage bewirkt eine Umsortierung der Ergebnisliste, sodass die Illusion, mit einem realen Suchsystem zu arbeiten, verstärkt wird (vgl. Abschn. 6.3.5). Sobald ein Dokument aus der Ergebnisliste selektiert wird, muss dieses, wie im ersten Experiment, zunächst bewertet werden, bevor weitere Treffer angesehen werden können. Es besteht zudem die Möglichkeit, Beurteilungen bereits bewerteter Dokumente im Nachhinein zu revidieren. Auch im zweiten Experiment wird bezüglich der Bearbeitungszeit lediglich eine maximale Grenze von zehn Minuten pro Aufgabe vorgegeben. Danach springt das Testsystem automatisch zurück zum Startbildschirm. Die Erhebung der Zufriedenheit der Testpersonen erfolgt jeweils im Anschluss an die Bearbeitung der Aufgaben. Als Grundlage dient in beiden Fällen der in Abschnitt 6.3.2 beschriebene Fragebogen, der nach Bearbeitung der zweiten Aufgabe um die in Abschnitt 6.3.1 beschriebenen Erwartungsitmes ergänzt wird. Erst danach werden die Testpersonen gebeten, den Wissenstest zur Erfassung ihres theoretischen und praktischen Suchmaschinenwissens zu bearbeiten, um weder das Suchverhalten noch die Zufriedenheitsreaktionen der Testpersonen zu beeinflussen. Abschließend wird in einer offenen Frage nach weiteren Anmerkungen zum Experiment gefragt und die in Abschnitt 6.3.1 erwähnte Kontrollfrage beantwortet. Als Anreiz zur Experimentteilnahme, haben alle Versuchspersonen darüber hinaus die Möglichkeit, an einem Wettbewerb teilzunehmen. Im Rahmen des Wettbewerbs kann jeder Teilnehmer für die beste Suchleistung einen von drei Geldpreisen im Wert von 50 €, 30 € bzw. 20 € gewinnen. Die Gesamtdauer des zweiten Experiment beläuft sich damit erneut auf circa 45 Minuten pro Teilnehmer.

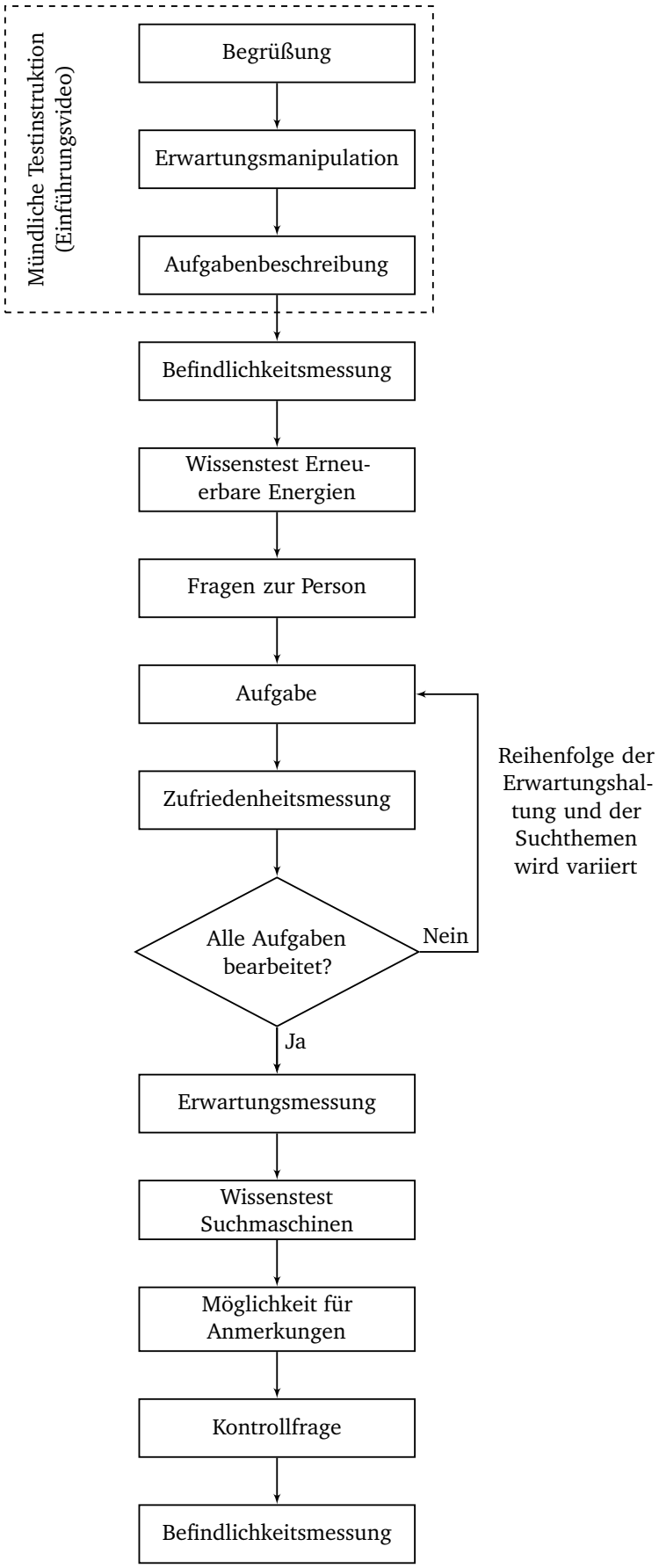


Abb. 6.10.: Schematische Darstellung des Versuchsablaufs des zweiten Experiments.

6.3.7. Ergebnisse des Pretests

Der qualitative Pretest gliedert sich in zwei Phasen. Im ersten Schritt wird ein Entwurf des in Abschnitt 6.3.2 beschriebenen Fragebogens einem Pretest durch Mitarbeiter des Instituts unterzogen, um die sprachliche und inhaltliche Verständlichkeit, die Vollständigkeit und die Relevanz der Fragen zu überprüfen. Dies erscheint insbesondere notwendig, da es sich bei den aus dem EUCS-Instrument übernommenen Items um eine Übersetzung der ursprünglich englischsprachigen Items handelt. Zusätzlich wird der Fragebogenentwurf in der zweiten Pretestphase auch Mitgliedern der Zielgruppe vorgelegt, um die Verständlichkeit der einzelnen Items sicherzustellen. Die Ergebnisse beider Phasen führen nur zu geringfügigen Änderungen in der Formulierung einzelner Items. Allerdings können nicht alle Änderungsvorschläge der Testpersonen bei der Überarbeitung des Fragebogens berücksichtigt werden. Dies gilt insbesondere für Vorschläge, die den Ausschluss von Items des EUCS-Instruments betreffen, da ihre Verwendung eine bessere Vergleichbarkeit mit anderen Untersuchungen gewährleisten soll. So wird bspw. die Relevanz einzelner Items als gering eingestuft. Auch wird eine inhaltliche Redundanz einzelner Items bemängelt. Jenseits dieser geäußerten Bedenken jedoch können die Fragen insgesamt als verständlich und der Aufbau des Fragebogens als logisch angesehen werden. Ein ähnliches Vorgehen wird auch im Fall der beiden Wissenstests angewandt. Auch sie können im Pretest als verständlich und nachvollziehbar bestätigt werden und werden daher unverändert in der Hauptuntersuchung verwendet.

Im zweiten Schritt wird der gesamte experimentelle Ablauf einer qualitativen Voruntersuchung durch zwei Mitglieder der adressierten Zielgruppe unterzogen. Dabei besteht das primäre Ziel dieser Pretestphase darin, die Funktionalität des Testsystems und die Qualität der Erwartungsmanipulation zu überprüfen. Neben der Methode des lauten Denkens werden die Versuchspersonen wie im ersten Experiment während der Aufgabenbearbeitung beobachtet, um einen zusätzlichen Eindruck von der Qualität des Untersuchungsmaterials zu erhalten. Die Ergebnisse der zweiten Pretestphase führen zu keinen wesentlichen Änderungen. In Bezug auf die Funktionalität des Testsystems fällt auf, dass ein Anklicken weiterführender Links innerhalb der gespeicherten Webdokumente zu einem Verlassen des Framesets des Testsystems führen kann. Aus diesem Grund werden nach dem Pretest noch einige für diesen Fehler besonders anfällige Seiten aus dem Testkorpus entfernt. Eine Überprüfung sämtlicher externer Links wird aus Zeitgründen jedoch nicht vorgenommen. Sowohl hinsichtlich der Verständlichkeit der Instruktionstexte und Fragebögen, als auch hinsichtlich der Gesamtdauer des Experiments ist das Feedback der Versuchspersonen positiv. Auch die Erwartungsmanipulation erweist sich als erfolgreich, denn die Kontrollfrage am Ende des Tests wird in beiden Fällen richtig beantwortet. Eine unumgängliche Schwierigkeit, die im Zusammenhang mit potentiellen Störvariablen (vgl. Abschn. 6.3.3) bereits thematisiert wird, betrifft den mit der freien Suchanfragenwahl verbundenen Kontrollverlust. Diesbezüglich geben beide Versuchspersonen im Nachhinein an, dass sie auch versucht hätten, von der Aufgabe abweichende Anfragen (auf Englisch oder zu einem Teilaspekt der Aufgabe) an die Suchmaschinen zu stellen und sich dann gewundert hätten, dass diese nicht erwartungsgemäß bedient werden. Die einzige Maßnahme, die gewährleistet, dass die Validität der Ergebnisse nicht durch die Formulierung der Suchanfragen beeinflusst wird und gleichzeitig genug Flexibilität bietet um eine realistischere Simulation des Suchprozesses zuzulassen, besteht in einer nachträglichen

Überprüfung der Suchanfragen, die im Rahmen der Hauptauswertung durchgeführt wird.

6.4. Ergebnisse

Die Hauptuntersuchung findet im Herbst 2009 an der Universität Hildesheim statt. In den folgenden Abschnitten werden die Ergebnisse des zweiten Experiments dargestellt. Der Ergebnisteil beginnt mit einer Beschreibung der Stichprobe nach demographischen Merkmalen. Diese mündet in der Erläuterung eines stichprobendifferenzierenden Auswertungskonzepts, das zum Ziel hat die Grenze zwischen Experiment und Anwendung auszuloten und sich das Spannungsgefüge zwischen mehr Kontrolle der Untersuchungsergebnisse auf der einen und mehr Realismus auf der anderen Seite zu eigen zu machen. Darauf folgt die hypothesenprüfende Auswertung der abhängigen Variablen. Während die Auswertung der Suchleistung äquivalent zum ersten Experiment anhand der in Abschnitt 6.3.2 beschriebenen Leistungsmaße erfolgt, wird vor der eigentlichen Auswertung der Benutzerzufriedenheit eine Faktorenanalyse zur Gruppierung der Zufriedenheitsitems durchgeführt. Wie im Fall des ersten Experiments, erfolgt in einem letzten Schritt eine Überprüfung des Einflusses verschiedener Faktoren, welche die Gütekriterien des Experiments möglicherweise einschränken.

6.4.1. Beschreibung der Stichprobe

Im Folgenden wird zunächst die untersuchte Stichprobe in ihrer Zusammensetzung beschrieben und mit dem ersten Experiment verglichen. Ein Teil der Tests kann im Rahmen einer regulären Lehrveranstaltung im Bereich der Informationswissenschaft durchgeführt werden. Die übrigen Teilnehmer werden über Mailinglisten, Aushänge und Direktansprache rekrutiert. Insgesamt nehmen 152 Personen an der Untersuchung teil. Im Rahmen der Stichprobenprüfung werden die Untersuchungsdaten anhand ihrer Eignung und ihres Gütegrads zunächst einer von drei Gütekategorien zugeteilt. Dabei wird unterschieden, ob die Daten ohne Einschränkung (Teilstichprobe SP_B), unter Vorbehalt (Teilstichprobe SP_{UV}) oder nicht (Teilstichprobe SP_N) in die Auswertung eingehen können (vgl. Tab. 6.3). Im Folgenden werden die einzelnen Kategorien eingehender erläutert.

Aufgrund technischer Probleme im Zusammenhang mit der Auswahl externer Links innerhalb der gespeicherten Webdokumente (vgl. Abschn. 6.3.7) können für 19 Teilnehmer keine vollständigen Daten erhoben werden. Diese Fälle werden von der Auswertung ausgeschlossen und die entsprechenden Dokumente am Ende des ersten Testtages im Testkorpus durch Dubletten anderer Dokumente ersetzt. Gleiches gilt für 11 weitere Teilnehmer, die entweder keine ausreichende Suchmotivation zeigen (z.B. nur ein Dokument innerhalb von 10 Minuten betrachten: 5 Fälle) oder durch die Eingabe themenfremder Suchbegriffe auffallen und somit den Wizard-of-Oz Charakter des Testsystems aufdecken (6 Fälle). Somit umfasst Kategorie SP_N insgesamt 30 Versuchspersonen.

In der Teilstichprobe SP_{UV} werden Testpersonen erfasst, bei denen es zu Auffälligkeiten im Rahmen der Testdurchführung kommt, die jedoch keinen Ausschluss aus der Untersuchung rechtfertigen. Diese lassen sich in drei Unterkategorien untergliedern, die im Folgenden erläutert werden. Dabei kann eine einzelne Testperson mehreren dieser Unterkategorien zugeordnet sein. Die größte, 56 Fälle umfassende Untergruppe betrifft Suchsessions, die entweder zu spezifische

Tab. 6.3.: Verteilung der Testteilnehmer auf die Untersuchungsgruppen ($n = 152$). Dargestellt ist die Gruppenverteilung in Abhängigkeit von den im Text beschriebenen Gütekategorien sowie dem Geschlecht (m/w) der Teilnehmer.

Erwartung		Stichprobe	System: A1 gut A2 gut			A1 gut A2 schlecht			A1 schlecht A2 gut			A1 schlecht A2 schlecht		
			m	w	ges.	m	w	ges.	m	w	ges.	m	w	ges.
A1 hoch A2 niedrig		SP_B	2	4	6	1	3	4	3	5	8	2	4	6
		SP_N	1	3	4	0	3	3	1	2	3	2	2	4
		SP_{UV}	1	9	10	2	8	10	1	6	7	0	8	8
		SP_{SB}	1	8	9	1	7	8	1	6	7	0	7	7
		SP_{MV}	0	1	1	0	0	0	0	0	0	0	1	1
		SP_{TD}	0	2	2	1	2	3	0	1	1	0	3	3
A1 niedrig A2 hoch		SP_B	0	3	3	1	7	8	3	9	12	4	4	8
		SP_N	1	3	4	2	2	4	2	3	5	0	3	3
		SP_{UV}	2	8	10	2	4	6	2	3	5	2	9	11
		SP_{SB}	2	6	8	2	4	6	1	2	3	2	6	8
		SP_{MV}	0	0	0	0	0	0	0	1	1	0	1	1
		SP_{TD}	0	2	2	0	1	1	1	1	2	0	3	3

oder zu allgemeine Suchbegriffe enthalten (Teilstichprobe SP_{SB}). Da die generischen Ergebnislisten des Testsystems auf solche Eingrenzungen bzw. Ausweitungen des Suchfokuses nicht adäquat reagieren können, ist nicht auszuschließen, dass dies zu einer Irritation der Probanden führen kann, weswegen diese Fälle im Rahmen der Auswertung gesondert berücksichtigt werden. Dabei ist die relativ hohe Anzahl dieser Fälle darauf zurückzuführen, dass hier im Sinne eines konservativen Ansatzes bereits eine einzelne auffällige Suchanfrage innerhalb einer Suchsession zur Aufnahme in die Kategorie SP_{SB} führt. Ein Beispiel für eine solche Suche ist der folgende Suchverlauf: *windenergie geräusche* > *windenergie lautstärke* > *windenergie geräuschentwicklung* > *windenergie aufbau* > *windenergie aufbau funktionsweise*. Während die ersten drei Sucheingaben das Thema vollständig erfassen, verschiebt sich der Fokus im Verlauf der Suche von der Geräuschentwicklung auf den Aufbau und die Funktionsweise von Windkraftanalagen, was im gegebenen Kontext als natürliches Suchverhalten gewertet werden kann.

Darüber hinaus wird in vier Fällen die Kontrollfrage zur Erwartungshaltung (vgl. Abschn. 6.3.1) nicht richtig beantwortet, was auf eine Versagen der Erwartungsmanipulation hinweist (Teilstichprobe SP_{MV}). Bei weiteren 17 Versuchspersonen deuten die Beantwortung der offenen Frage oder ein informelles Gespräch im Anschluss an das Experiment darauf hin, dass der Versuchsaufbau möglicherweise durchschaut wurde (Teilstichprobe SP_{TD}). Da in diesen Fällen nicht festgestellt werden kann, ob die in der Testinstruktion enthaltene Erwartungsmanipulation trotzdem unbewusst verarbeitet wird, verbleiben auch diese Datensätze vorerst in der Stichprobe und werden im Rahmen der Auswertung gesondert betrachtet.

Tabelle 6.3 gibt die Verteilung der Probanden aufgeschlüsselt nach Auswertbarkeit und Geschlecht auf die acht Versuchsgruppen wieder. Dabei entspricht SP_B allen Fällen, bei denen keinerlei Probleme auftreten, SP_N umfasst Fälle, die aus oben genannten Gründen von der Auswertung ausgeschlossen werden müssen und SP_{UV} beinhaltet diejenigen Fälle, bei denen die aufgetretenen Unregelmäßigkeiten keinen direkten Ausschluss rechtfertigen (SP_{SB}, SP_{MV} u. SP_{TD}). Da einige Probanden innerhalb der UV-Gruppe mehreren Kategorien zugeordnet werden können, kann es vorkommen, dass die Gesamtanzahl dieser Oberkategorie in einigen Fällen niedriger

Tab. 6.4.: Demographische Daten. Nach der Datenbereinigung liegt eine Stichprobe von $n = 122$ Untersuchungsteilnehmern vor, was einer Stichprobenausschöpfung von $80\,\%$ entspricht.

Variable	**Maß**	**Wert**	
Alter	Median	22	
	Standardabweichung	3	
	Spanne	17 – 38	
	Mittelwert	23	
	Kategorie	**Anzahl**	**Prozent**
Muttersprache	Deutsch	108	88,5
	zweisprachig	3	2,5
	nicht Deutsch	11	9,0
Geschlecht	männlich	27	22,1
	weiblich	95	77,9
Tätigkeit	Student	115	94,3
	Doktorand	5	4,1
	Mitarbeiter	1	0,8
	Schüler	1	0,8
Fachbereich	Erziehungs- und Sozialwissenschaften	50	43,5
	Kulturwissenschaften und ästhetische Kommunikation	8	7,0
	Sprach- und Informationswissenschaften	43	37,4
	Mathematik, Naturwissenschaften, Wirtschaft und Informatik	13	11,3
	Sonstige	1	0,9

ausfällt als die Summe ihrer Unterkategorien.

In Tabelle 6.4 sind die demographischen Merkmale der 122 in der Stichprobe verbleibenden Versuchspersonen aufgeführt (Kategorien SP_B u. SP_{UV}). Die Stichprobe umfasst somit 27 Männer und 95 Frauen im Alter von 17 bis 38 Jahren. Das Medianalter liegt bei 22 Jahren. Im Schnitt sind Teilnehmer des zweiten Experiments damit etwas jünger als Teilnehmer des ersten Experiments. Zwar kommt es auch im zweiten Experiment zu einer geringfügige Abweichung von der ursprünglich angestrebten Altersspanne, da Alterseffekte jedoch im ersten Experiment nur eine geringe Rolle spielen und zusätzlich im Rahmen einer Kovarianzanalyse überprüft werden, kann diese Tatsache als unproblematisch angesehen werden. Der Anteil der Probanden nicht ausschließlich deutscher Muttersprache ist vergleichbar mit dem des ersten Experiments. Auf die Frage, ob Deutsch ihre Muttersprache ist, antworten $89\,\%$ der Probanden mit ja, $2\,\%$ geben an zweisprachig aufgewachsen zu sein und $9\,\%$ haben eine andere Muttersprache als Deutsch. Damit beträgt der Anteil der Probanden nichtdeutscher Muttersprache $11\,\%$ im Vergleich zu $13\,\%$ im ersten Experiment. Studierende stellen mit $94\,\%$ auch im zweiten Experiment den Hauptteil der Probanden dar. Tabelle 6.4 verdeutlicht darüber hinaus, wie sich die Studierenden auf die an der Universität Hildesheim gelehrten Fachbereiche verteilen. Dabei ist zu beachten, dass alle lehramtsbezogenen Studiengänge in dieser Statistik dem Fachbereich Erziehungs- und Sozialwissenschaften zugeordnet sind. Bis auf fünf Teilnehmer aus dem dritten, fünften und sechsten Semester sind die übrigen 29 Versuchspersonen aus IT-orientierten Studiengängen Studierende im ersten Semester. Da auch der Einfluss der Sucherfahrung auf die Untersuchungsergebnisse überprüft wird, kann auch dies als unproblematisch angesehen werden.

Tabelle 6.5 fasst die auf die Computer- und Sucherfahrung bezogenen Merkmale der Stichprobe zusammen. Die durchschnittliche Computererfahrung (M) liegt bei 9,1 Jahren (SD: 3,2), die durchschnittliche Suchmaschinenerfahrung beträgt 6,7 Jahre (SD: 2,4). Hinsichtlich der Computernutzung geben zehn Teilnehmer an, über mindestens 15 Jahre Computererfahrung zu verfügen. Bezieht man das Alter der Probanden mit ein, zeigt sich, dass das früheste Alter der

Tab. 6.5.: Computer- und Sucherfahrung (n = 122).

Variable	Median	M	SD	Spanne
Computererfahrung (Jahre)	9	9,1	3,2	1 - 20
Computernutzung (Stunden pro Woche)	16	18,7	11,6	2 - 75
Suchmaschinenerfahrung (Jahre) (n = 121)	6	6,7	2,4	2 - 15
Suchmaschinennutzung (Stunden pro Woche) (n = 119)	2	4,3	5,2	0,5 - 30
Bekannte Suchmaschinen	2	2,5	1,3	1 - 7
Verwendete Suchmaschinen	1	1,3	0,5	1 - 3

Computernutzung bei sechs Jahren (also im Schuleintrittsalter) liegt.

Hinsichtlich der sowohl im ersten als auch im zweiten Experiment erfassten Merkmale wöchentliche Computernutzung, Bekanntheit und Nutzung unterschiedlicher Suchdienste ergibt sich ein ähnliches Bild. Eine Vergleichbarkeit der Stichproben ist somit gegeben. Die durchschnittliche Computernutzung der Probanden liegt mit $18{,}7$ Stunden pro Woche (SD: $11{,}6$) nur leicht über den im ersten Experiment erhobenen durchschnittlichen Nutzungswerten (M: $16{,}7$; SD: $12{,}8$). Wie schon im ersten Experiment beobachtet, fällt der Unterschied zwischen Minimum und Maximum recht groß aus. Als mögliche Ursache wird erneut die Abhängigkeit der Nutzungsintensität von der aktuellen Studienphase angenommen. In Bezug auf die Bekanntheit und die Nutzung unterschiedlicher Suchdienste zeigt sich in dieser Stichprobe, dass die Befragten im Schnitt $2{,}5$ unterschiedliche Suchmaschinen kennen (SD: $1{,}3$), jedoch nur $1{,}3$ Suchmaschinen regelmäßig verwenden (SD: $0{,}5$). Die geringen Werte für die Standardabweichungen deuten wie im ersten Experiment darauf hin, dass in diesem Fall von einem vergleichbaren Wissensstand ausgegangen werden kann. Anstelle der Internetnutzungsfrequenz wird im zweiten Experiment die Suchmaschinennutzungsfrequenz erhoben. Für die 119 Probanden, die diese Frage beantwortet haben, liegt die durchschnittliche Suchmaschinennutzung bei $4{,}3$ Stunden pro Woche (SD: $5{,}2$). Die Spanne der angegebenen Zeitintervalle fällt mit einer minimalen Nutzungsdauer von $0{,}5$ und einer maximalen Nutzungsdauer von 30 Stunden pro Woche erneut relativ groß aus, sodass nicht auszuschließen ist, dass Ausreißer die Ergebnisse beeinflussen.

Tab. 6.6.: Selbsteinschätzung des Domänen- und Suchmaschinenwissens (n = 122).

	Domänenwissen		Suchwissen	
Kategorie	Anzahl	Prozent	Anzahl	Prozent
Anfänger	100	82,0	52	42,6
Fortgeschrittene	22	18,0	68	55,8
Experten	0	0	2	1,6

Die Ergebnisse zur Selbsteinschätzung der Probanden zeigen, dass die Teilnehmer in beiden Bereichen über ein laienhaftes bis fortgeschrittenes Wissen verfügen (vgl. Tab. 6.6). Während die Teilnehmer ihre Kenntnisse in Bezug auf die Suchdomäne mehrheitlich auf Anfängerniveau einstufen ($82{,}0\,\%$), verteilen sich die Selbsteinschätzungen hinsichtlich ihrer Suchmaschinenkenntnisse nahezu in gleichem Maße auf die Kategorien Anfänger ($42{,}6\,\%$) und Fortgeschrittene ($55{,}7\,\%$). Die Gruppe der Experten ist gemäß dieser Selbsteinschätzung in der Stichprobe fast nicht vertreten. Nur zwei Versuchsteilnehmer stufen sich als Experten hinsichtlich ihrer Suchmaschinenkompetenz ein.

Die Ergebnisse der beiden Wissenstests sind in Abbildung 6.11 dargestellt. Diese scheinen auf

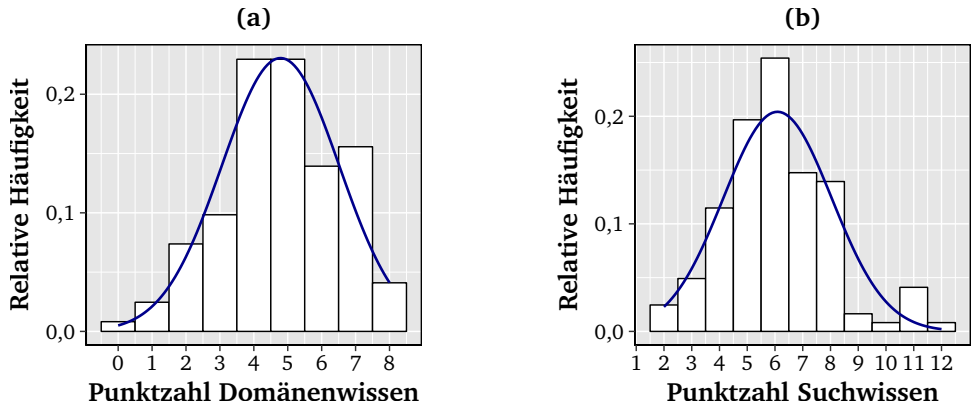


Abb. 6.11.: Prozentuale Verteilung der Testergebnisse der Wissenstests zum Suchmaschinen- und Domänenwissen. Bild (a): Prozentuale Verteilung der Testergebnisse des Wissenstests zur Suchdomäne nach Punktzahl. Bild (b): Prozentuale Verteilung der Testergebnisse des Wissenstests zu Suchmaschinen nach Punktzahl.

eine Normalverteilung hinzudeuten. In beiden Fällen erstrecken sich die Werte der Gesamtpunktzahl über die gesamte Bandbreite der Skala. Dies zeigt, dass die Skala Unterschiede zwischen den Probanden gut erfasst. Auch die Mittelwerte liegen in beiden Fällen etwa in der Mitte der Skala. Die durchschnittlich erreichte Punktzahl liegt für den Wissenstest zur Suchdomäne bei $4{,}8$ von insgesamt acht möglichen Punkten (SD: $1{,}7$), beim Suchmaschinenwissen beträgt die durchschnittlich erreichte Punktzahl $6{,}1$ von insgesamt zwölf möglichen Punkten (SD: $2{,}0$). Eine Überprüfung der Normalverteilungshypothese mit dem Shapiro-Wilk-Test ergibt jedoch in beiden Fällen, dass die Ergebnisse der Wissenstests nicht normalverteilt sind. Eine graphische Analyse der Daten lässt jedoch auf eine um den Mittelwert symmetrische Verteilung schließen.

Tab. 6.7.: Ergebnisse des Wissenstests zum Domänen- und Suchmaschinenwissen (n = 122). Zum Vergleich von Selbsteinschätzung und Wissenstests wurden die Punktwerte der Wissenstestskalen zu drei Kategorien zusammengefasst. Im Fall des Domänenwissens wird wie folgt umkodiert: 0-2 Punkte = Anfänger, 3-5 Punkte = Fortgeschrittene, 6-8 Punkte = Experten. Im Fall des Suchwissens fallen die Intervalle entsprechend der breiteren Skala folgendermaßen aus: 1-4 Punkte = Anfänger, 5-8 Punkte = Fortgeschrittene, 9-12 Punkte = Experten.

	Domänenwissen		Suchwissen	
Kategorie	**Anzahl**	**Prozent**	**Anzahl**	**Prozent**
Anfänger	13	10,7	23	18,9
Fortgeschrittene	68	55,7	90	73,8
Experten	41	33,6	9	7,4

Im Vergleich zur Selbsteinschätzung zeigt sich, dass die Probanden dazu neigen ihren eigenen Wissensstand zu unterschätzen. Dies wird insbesondere deutlich, wenn man die beiden Skalen in eine dreistufige Skala umkodiert (vgl. Tab. 6.7). Während die Teilnehmer bei der Selbsteinschätzung ihre Kenntnisse bezüglich der Suchdomäne mehrheitlich auf Anfängerniveau einstufen ($82{,}0\,\%$), ergibt der Wissenstest, dass die meisten Teilnehmer tatsächlich über ein deutlich höheres Wissen verfügen (Fortgeschrittene: $55{,}7\,\%$; Experten: $33{,}6\,\%$). Auch in Bezug auf die Suchmaschinenkenntnisse der Testpersonen ist im Vergleich eine leichte Verschiebung der Verteilung erkennbar. So stufen $40{,}7\,\%$ der Versuchspersonen ihr Wissen über Suchmaschinen auf Anfängerniveau ein. Gemäß den Ergebnissen des Wissenstests verfügen jedoch nur $18{,}9\,\%$ der

Teilnehmer über ein eher geringes Wissen. Um Zusammenhänge zwischen der Selbsteinschätzung der Domänen- und Suchkompetenz und den tatsächlichen Kenntnissen der Experimentteilnehmer statistisch zu überprüfen, wird außerdem der Spearman'sche Rangkorrelationskoeffizient errechnet. Es findet sich in beiden Fällen ein schwacher bis mäßiger Zusammenhang ($p < 0{,}01$). Für das Domänenwissen beträgt der Korrelationskoeffizient $r = 0{,}34$ und für das Suchmaschinenwissen $r = 0{,}28$.

Zusammenfassend lässt sich festhalten, dass die für die Auswahl der Stichprobe des zweiten Experiments gesteckten Ziele erreicht werden. Die vorliegende Stichprobe lässt sich hinsichtlich der meisten soziodemographischen Kriterien als relativ homogen bezeichnen. Kritisch kann der hohe Anteil weiblicher Probanden bewertet werden, wobei dieser jedoch das vorherrschende Verhältnis weiblicher und männlicher Studierender an der Universität Hildesheim widerspiegelt. Der Vergleich mit der Stichprobe des ersten Experiments ergibt darüber hinaus, dass die Teilnehmer beider Experimente aus einer vergleichbaren Grundgesamtheit stammen.

6.4.2. Auswertungskonzept

Im Rahmen von Untersuchungen zum interaktiven Information Retrieval bewegt man sich zwangsläufig im Spannungsfeld zwischen Experiment und Anwendungsfall. Je mehr die Rahmenbedingungen eines Experiments kontrolliert und damit Störeinflüsse ausgeblendet werden, desto größer ist die Gefahr, sich von der realen Anwendungssituation zu entfernen. Umgekehrt sind die spezifischen Effekte, die im Rahmen der Untersuchung adressiert werden sollen, in einem zu offenen Versuchsaufbau unter Umständen nicht eindeutig nachweisbar. Ähnliche Fragen stellen sich auch im Zuge der Datenaufbereitung, da sich, wie in Abschnitt 6.4.1 dargestellt, aus Nachgesprächen, Antworten auf offene Fragen oder den verwendeten Suchbegriffen, Ausschlussgründe für bestimmte Versuchsteilnehmer ergeben können.

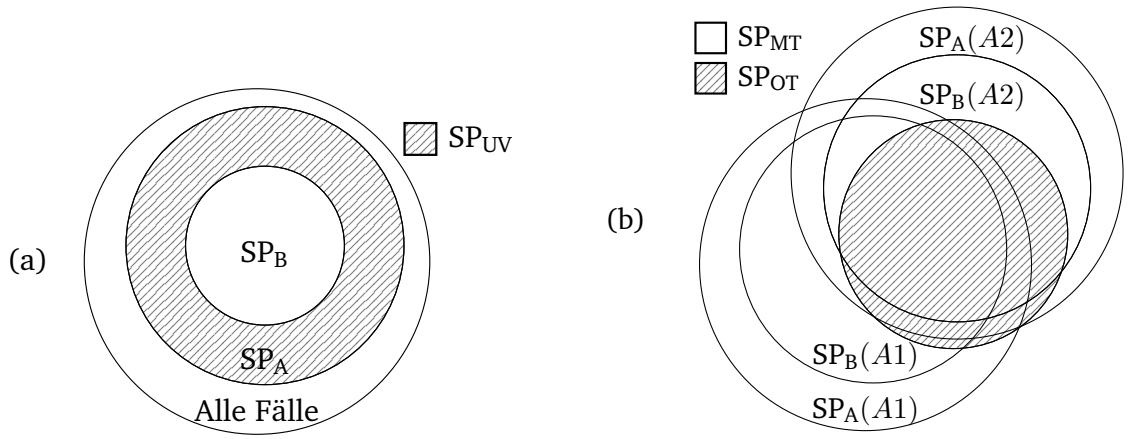


Abb. 6.12.: Schematische Darstellung über die Relationen zwischen den Teilstichproben SP_A und SP_B. Bild (a) zeigt das Verhältnis der Teilstichproben SP_A, SP_{UV} und SP_B. Bild (b) skizziert exemplarisch wie die Gruppe der eindeutig keinen Topiceffekt aufweisenden Variablen (SP_{OT}) im Verhältnis zu den Stichproben SP_A und SP_B der beiden Aufgaben A_1 und A_2 aufgefasst werden kann.

Vor diesem Hintergrund wird zur Auswertung des zweiten Experiments ein stichprobendifferenzierendes Auswertungskonzept entwickelt, das die Auswertung kritischer Fälle nicht von vornherein ausschließt. Ausgehend von den in Tabelle 6.3 beschriebenen Untermengen SP_B, SP_{UV}

und SP_N besteht die Möglichkeit, den Realitätsgrad der Untersuchung anzupassen. Dazu wird entweder der vollständig bereinigte Datensatz SP_B oder der weniger streng kontrollierte Gesamtdatensatz SP_A, bestehend aus allen Fällen der Teilstichproben SP_B und SP_{UV}, für die Auswertung herangezogen. Abbildung 6.12 Bild (a) veranschaulicht noch einmal graphisch, dass sich beide Stichproben gerade um die Datensätze in Teilstichprobe SP_{UV} unterscheiden. Entsprechend obiger Argumentation wird davon ausgegangen, dass Effekte die sich in beiden Datensätzen nachweisen lassen, eine größere Generalisierbarkeit besitzen sollten.

Sowohl die nachträgliche Bereinigung der Stichprobe als auch fehlende Werte in den abhängigen Variablen können im zweiten Experiment zu einer ungleichmäßigen Verteilung der beiden Suchthemen beitragen. Dabei kann das Fehlen von Werten verschiedene Gründe haben: So kann es z.B. im Kontext der Benutzerzufriedenheit vorkommen, dass Probanden einzelne Fragen nicht beantworten, was sich im Umkehrschluss auf die Berechenbarkeit einzelner Zufriedenheitsskalen auswirken kann. Ebenso können einzelne Leistungsmaße, die als Quotient zweier Leistungsindikatoren definiert sind, nicht berechnet werden, da in diesen Fällen durch 0 geteilt werden müsste (vgl. Abschn.6.3.2). Aus diesem Grund werden im nächsten Schritt pro Datensatz, Aufgabe und abhängiger Variable je fünf nach Versuchsgruppe und Suchthema ausbalancierte Zufallsstichproben gezogen, um Topic- und Reihenfolgeeffekte zu minimieren. Durch die Auswahl mehrerer Teilstichproben soll außerdem verhindert werden, dass zufällige Auswahleffekte die Untersuchungsergebnisse verfälschen.

Um jeweils die größtmögliche Fallzahl berücksichtigen zu können, wird eine weitere Qualitätsstufe eingeführt, die unterscheidet, ob für eine gegebene Variable ein Topiceffekt vorliegt, also signifikante Unterschiede zwischen den einzelnen Suchaufgaben bestehen (vgl. Abschn. 4.2.3.2). Liegt für ein Benuzterleistungs- bzw. Zufriedenheitsmaß ein solcher Aufgabeneffekt vor, erfolgt die Balancierung der beiden Datensätze wie oben beschrieben sowohl nach Versuchsgruppe als auch nach Suchthema (SP_{MT}). Kann hingegen eindeutig nachgewiesen werden, dass für eine Variable kein Topiceffekt vorhanden ist, wird ein zusätzlicher Datensatz ohne Topic-Balancierung angelegt (SP_{OT}). Auch in diesem Fall werden wiederum je fünf Zufallsstichproben gezogen. Um das Vorhandensein von Topiceffekten zwischen den beiden Suchthemen zu überprüfen, werden einfaktorielle Varianzanalysen mit den Leistungs- und Zufriedenheitsmaßen als abhängige und mit dem Suchthema als unabhängige Variable durchgeführt (vgl. Abschn. 6.4.6.1). Dabei gilt in dieser Arbeit das Vorhandensein eines Topiceffekts für eine abhängige Variable als ausgeschlossen, wenn jeweils für alle fünf Zufallsstichproben von SP_B und SP_A weder für die erste noch für die zweite Aufgabe ein signifikanter Einfluss des Suchthemas zu beobachten ist. Das Venn-Diagramm in Abbildung 6.12 Bild (b) stellt die beiden Mengen SP_{MT} und SP_{OT} noch einmal graphisch dar. Damit wird deutlich, dass die Menge SP_{OT} gerade die abhängigen Variablen enthält, die eindeutig keine Aufgabenabhängigkeit zeigen, während SP_{MT} die Variablen umfasst, für die entweder eindeutig ein Topiceffekt nachgewiesen oder nicht eindeutig ausgeschlossen werden kann. Eine Übersicht über diejenigen Variablen, die nach obiger Definition eindeutig keinen Topiceffekt aufweisen, ist außerdem in Tabelle C.25 in Anhang C.4 zu finden.

Als weiterer Punkt ist an dieser Stelle die Mindeststichprobengröße zu erwähnen, die in Abhängigkeit von Qualitätsstufe und Versuchsgruppenanzahl angepasst wird. Dabei ist neben den statistischen Gütekriterien auch die vorhandene Datenlage zu berücksichtigen. So fällt die Anzahl

der zur Verfügung stehenden Datensätze in der bereinigten Stichprobe SP_B notwendigerweise geringer aus als dies für den weniger kontrollierten Datensatz SP_A der Fall ist. Weiterhin ist zu beachten, dass sich die Testpersonen für die mit dem ersten System bearbeite Aufgabe auf lediglich vier unterschiedliche Untersuchungsgruppen verteilen (System 1 vs. Erwartung). Für die als zweites bearbeitete Aufgabe sind hingegen acht unterschiedliche Treatmentgruppen zu berücksichtigen (System 1 vs. System 2 vs. Erwartung), was mit einer Reduzierung der Stichprobengröße pro Versuchsgruppe einhergeht. Im Rahmen der Auswertung werden daher in Abhängigkeit der vorhandenen Datenbasis unterschiedlich große Mindeststichprobengrößen zugrunde gelegt. Genauer wird für SP_A die Stichprobengröße für die erste Aufgabe mit vier unterschiedlichen Versuchsgruppen auf mindestens 20 Fälle pro Gruppe festgelegt, für die zweite Aufgabe mit acht unterschiedlichen Versuchsgruppen hingegen auf mindestens 10 Fälle pro Gruppe. Da SP_B insgesamt nur 110 Fälle beinhaltet, wird die Mindeststichprobengröße vor dem Hintergrund der homogeneren Stichprobenzusammensetzung für beide Aufgaben eine Mindestanzahl von 10 Fälle pro Gruppe akzeptiert. Es wird davon ausgegangen, dass sich bei einer Teilnehmerzahl von weniger als 10 Personen pro Gruppe keine statistisch validen Ergebnisse mehr ableiten lassen. Dies führt dazu, dass für die Menge SP_B nur für die erste Aufgabe und nur für Variablen, bei denen ein Topiceffekt ausgeschlossen werden kann, genügend Daten zur Verfügung stehen. Tabelle 6.8 enthält eine Übersicht über die vorhandenen Datensätze sowie die darin enthaltenen Variablen. Pro Datensatz werden fünf Zufallsstichproben gezogen, sodass pro Variable maximal 25 Stichproben zu vergleichen sind.

Tab. 6.8.: Übersicht über verfügbare Datensätze. Neben der Anzahl der unabhängigen Variablen mit ausreichender Fallzahl (Varanz.), sind auch die minimale (n_{min}) sowie die mittlere (n_{mean}) Stichprobengröße angegeben.

		Benutzerleistung			**Zufriedenheit**		
Aufgabe	**Datensatz**	**Varanz.**	**n_{min}**	**n_{mean}**	**Varanz.**	**n_{min}**	**n_{mean}**
$A1$	$SP_{A,MT}$	68	88	95	56	96	96
$A1$	$SP_{A,OT}$	21	80	103	9	108	108
$A1$	$SP_{B,OT}$	17	40	40	6	40	40
$A2$	$SP_{A,MT}$	60	80	80	56	80	80
$A2$	$SP_{A,OT}$	17	96	96	9	96	96

Um die Darstellung der Ergebnisse so übersichtlich wie möglich zu gestalten, wird die in den weiteren Abschnitten folgende Darstellung der Ergebnisse vom Allgemeinen zum Speziellen verlaufen. Konkret bedeutet dies, dass zunächst nur Ergebnisse der Menge SP_A berichtet werden und erst im Anschluss diskutiert wird, ob der Ausschluss kritischer Fälle das Untersuchungsergebnis signifikant verändert.

Als Fazit lässt sich festhalten, dass die Auswertung kritischer Fälle durch das vorgestellte Konzept nicht von vornherein abgelehnt wird. Vielmehr wird mit dem Ziel einer optimalen Ausnutzung der Daten von einer strengen Unterscheidung in wertbare und nicht wertbare Fälle abgesehen und stattdessen eine Hierarchie unterschiedlicher Qualitätsstufen eingeführt. Dieses Vorgehen erlaubt es, die Stabilität der beobachteten Effekte in Bezug auf die Hinzunahme kritischer Fallgruppen zu untersuchen und somit Rückschlüsse auf die Zuverlässigkeit bzw. externe Validität der jeweiligen Befunde zu erhalten.

6.4.3. Auswertung der Benutzerleistung

In diesem Abschnitt wird die Suchleistung der Untersuchungsteilnehmer anhand der in Abschnitt 6.3.2 beschriebenen Leistungsmaße analysiert. Analog zum ersten Experiment bilden Systemleistung und Benutzererwartung die unabhängigen Variablen, die zur Erklärung der abhängigen Variable Benutzerleistung herangezogen werden. Aufgrund der Tatsache, dass das zweite Experiment die Nutzung zweier vermeintlich unterschiedlicher Systeme vorsieht (vgl. Abschn. 6.3), erfordert die Auswertung zwei separate Varianzanalysen. Dabei sind prinzipiell zwei Möglichkeiten denkbar: Zum einen ist es möglich, beide Systeme unabhängig voneinander zu betrachten und das Risiko in Kauf zu nehmen, dass die Nutzungserfahrung des ersten Systems einen zusätzlichen, konfundierenden Einfluss auf die Ergebnisse des zweiten Systems besitzt. Die zweite Möglichkeit besteht darin, die erste Aufgabe analog zum ersten Experiment mittels zweifaktorieller Varianzanalyse zu untersuchen, die Nutzungserfahrung des ersten Systems aber bei der zweiten Aufgabe als weiteren Faktor in die Analyse einzubeziehen. Da im Rahmen dieser Arbeit davon ausgegangen wird, dass bereits gemachte Erfahrungen einen Einfluss auf das zukünftige Suchverhalten und die Wahrnehmung von Suchergebnissen haben können, wird die zweite Möglichkeit gewählt, sodass im Fall der zweiten Aufgabe dreifaktorielle Varianzanalysen mit den Faktoren Benutzererwartung sowie Systemleistung des ersten und zweiten Systems durchgeführt werden. Im Gegensatz zur Systemleistung genügt es, die Benutzererwartung einmal in die Analyse einzubeziehen, da eine niedrige Erwartungshaltung im Rahmen der ersten Aufgabe automatisch eine hohe Erwartungshaltung für das zweite präsentierte System impliziert. Wie bereits im Zusammenhang mit der Vorstellung des Auswertungskonzepts erwähnt (vgl. Abschn. 6.4.2), führt dies dazu, dass die verfügbare Anzahl an Testpersonen pro Versuchsgruppe sich im Fall der zweiten Aufgabe auf etwa die Hälfte reduziert.

Der Umfang der betrachteten abhängigen Variablen macht es notwendig, von einer ausführlichen Diskussion jedes einzelnen Benutzerleistungsmaßes abzusehen. Vielmehr konzentriert sich die folgende Darstellung darauf, Beobachtungen, die anhand verschiedener Indikatoren belegt werden können, zu bündeln. Dabei erfolgt die Argumentation häufig anhand einiger weniger ausgewählter Leistungsmaße, während weitere, die jeweilige These stützende, Variablen durch Angabe in Klammern dokumentiert werden. Dies ermöglicht einerseits, das analytische Vorgehen nachzuvollziehen und andererseits, die den Befunden zugrunde liegende Datenbasis zu protokollieren. In diesem Sinne können diese Auflistungen als weiteres Maß der Stabilität bzw. Stärke der jeweiligen Effekte interpretiert werden, die zum inhaltlichen Verständnis der Argumentationslinien jedoch nicht notwendigerweise in ihrer Gänze nachvollzogen werden müssen.

Berichtet werden im Folgenden ausschließlich Ergebnisse, für die ein signifikanter Zusammenhang zwischen einer der unabhängigen Variablen und dem untersuchten Leistungsmaß eindeutig (d.h. über alle fünf Zufallsstichproben hinweg) oder in der Tendenz (d.h. für vier der fünf Zufallsstichproben) nachgewiesen werden kann. Nicht berichtet werden hingegen weniger stabile Ergebnisse, die nur in ein, zwei oder drei der Zufallsstichproben auftreten. Die Tabellen 6.9 und 6.10 enthalten die Gruppenmittelwerte der Benutzerleistungsmaße mit eindeutig oder in der Tendenz signifikanten Effekten. Um die Präsentation der Ergebnisse so übersichtlich und platzsparend wie möglich zu gestalten, wird pro Variable nur der Werteverlauf der Stichprobe mit dem signifikantesten Ergebnis exemplarisch dargestellt. Eindeutig sowie tendenziell signifikante

Tab. 6.9.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung für $A1$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen C.17 und C.23 in Anhang C.3 entnommen werden.

		System		Erwartung	
ID	**Beschreibung**	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$
M05	Anz. aufg. irrel. Dok.	**2,48**[a]	**5,90**	4,63	3,75
M08	Anz. falsch rel. bew. Dok.	**0,81**[a]	**1,90**	1,59	1,12
M15	Anz. richtig irrel. bew. Dok.	**1,69**	**3,78**[a]	2,80	2,67
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	**0,71**	**0,81**[a]	0,76	0,76
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,21**[a]	**0,40**	0,30	0,31
V02	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,79**[a]	**0,61**	0,70	0,71
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,25**	**0,15**[a]	0,20	0,19
V06	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	**0,57**	**0,32**[a]	0,49	0,40
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,07**[a]	**0,15**	0,13	0,09
V09	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	**0,12**[a]	**0,24**	0,19	0,17
V10	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. richtig rel. bew. Dok.}}$	**0,15**[a]	**0,40**	0,30	0,25
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,15**	**0,24**[a]	0,18	0,21
V17	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	**0,43**	**0,69**[a]	0,52	0,60
V19	$\frac{\text{Anz. richtig rel. bew. Dok. (erste 10 Dok.)}}{\text{Anz. rel. bew. Dok. (erste 10 Dok.)}}$	**0,90**[a]	**0,75**	0,82	0,83
V21	$\frac{\text{Anz. richtig rel. bew. Dok. (erste Suche)}}{\text{Anz. rel. bew. Dok. (erste Suche)}}$	**0,90**[a]	**0,81**	0,84	0,88
V25	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. rel. bew. Dok. (letzte Suche)}}$	**0,90**[a]	**0,79**	0,85	0,84
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,56**[a]	**0,47**	0,51	0,52
V31/BP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	**0,91**[a]	**0,78**	0,83	0,87

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Effekte sind in den Tabellen fett hervorgehoben. Darüber hinaus bleibt zu erwähnen, dass die Ergebnisse je nach Erfüllung der Verteilungsvoraussetzungen auf einer klassischen oder einer robusten Varianzanalyse beruhen (vgl. Abschn. 4.3.2). Weitergehende Informationen, wie das im Einzelnen verwendete Verfahren, das Vorliegen eines tendenziellen bzw. eindeutigen Effekts und das genaue Signifikanzniveau, können hingegen den Tabellen C.23 und C.24 in Anhang C.3 entnommen werden. Wie bereits in Abschnitt 6.4.2 dargelegt, stehen für die bereinigte Stichprobe SP_B nur für die erste Aufgabe und nur für Variablen, bei denen ein Topiceffekt eindeutig auszuschließen ist, ausreichend Daten zur Verfügung, um den Mindeststichprobenumfang von 10 Probanden pro Untersuchungsgruppe zu erreichen. Eindeutig oder in der Tendenz signifikante Unterschiede sind mit der Ausnahme des Leistungsmaßes V02 bei der ersten Aufgabe nur für die

Gesamtstichprobe SP_A beobachtbar. In zwölf Fällen ergeben sich in SP_B eindeutig keine signifikanten Effekte, was in elf Fällen das Ergebnis für SP_A bestätigt. Einzig der für das Leistungsmaß M15 in SP_A signifikante Systemeinfluss ist für SP_B in keiner der fünf Stichproben nachweisbar. Bei den verbleibenden vier Leistungsmaßen aus der Stichprobe SP_B bei denen über die fünf Zufallsstichproben weder eine Entscheidung zugunsten eines Effektes noch seiner Abwesenheit getroffen werden kann, wird in einem Fall das Ergebnis aus SP_A bestätigt (M18), während in zwei Fällen (M01 u. S05) in SP_A eindeutig kein Effekt vorliegt bzw. für V14 ein eindeutiger Systemeffekt nachweisbar ist (vgl. Tab. 6.9). Abweichungen zwischen SP_A und SP_B treten damit insgesamt nur für wenige Fälle auf. Daher basiert die Auswertung wenn nicht anders angegeben auf der Stichprobe SP_A. Hier beträgt der Gesamtstichprobenumfang sowohl für Aufgabe 1 als auch für Aufgabe 2 mindestens 80 Versuchspersonen (vgl. Tab. 6.8).

Von den 86 betrachteten Leistungsvariablen zeigen bei der ersten Aufgabe 28 und bei der zweiten Aufgabe 23 Variablen einen signifikanten Unterschied in Bezug auf den Einfluss der Systemleistung. Demgegenüber scheint die Erwartungshaltung eine deutlich geringere Rolle bei der Bewertung der Suchergebnisse zu spielen. Nur für eine einzige Variable (V28/PCP) zeigt sich bei der zweiten Aufgabe in der Tendenz ein signifikanter Zusammenhang zwischen der Erwartungshaltung der Benutzer und ihrer Suchleistung.

Die meisten Unterschiede in Bezug auf die Systemleistung sind über alle fünf Stichproben hinweg sichtbar. Lediglich in drei Fällen der ersten und einem Fall der zweiten Aufgabe handelt es sich um tendenzielle Effekte. Dreizehn der Variablen, die bei der ersten Aufgabe signifikant werden, zeigen auch bei der zweiten Aufgabe einen signifikanten Einfluss, sodass insgesamt nur fünf Effekte nicht bestätigt werden können. Ein Blick auf die betreffenden Variablen zeigt überdies, dass bei der zweiten Aufgabe in drei Fällen keine Varianzanalyse durchgeführt werden kann, weil nicht genügend Probanden vorhanden sind ($n < 80$). In den übrigen beiden Fällen hingegen ist das Ergebnis nicht aussagekräftig: Der Effekt kann weder eindeutig nachgewiesen, noch ausgeschlossen werden. Der Fall, dass ein für die erste Aufgabe vorhandener Effekt bei der zweiten Aufgabe über alle Stichproben hinweg nicht signifikant ist, tritt also nicht auf. Eine weitere interessante Beobachtung betrifft diejenigen Variablen, für die nur im Kontext der zweiten Aufgabe signifikante Unterschiede zwischen den Versuchsgruppen festgestellt werden können. Hier zeigt sich, dass für acht der zehn Variablen (M03, M12, M11, M17, M18, S01, S05, S05-log) bei der ersten Aufgabe über alle fünf Stichproben hinweg kein signifikanter Unterschied in der Benutzerleistung besteht (vgl. Anh. C.3, Tab. C.16). Dies legt die Vermutung nahe, dass diese Variablen Aspekte der Benutzerleistung erfassen, die einer gewissen Einarbeitung bedürfen. Besonders einleuchtend erscheint dies bspw. bei S05, der Zeit bis zum ersten richtig relevant bewerteten Dokument.

Des Weiteren lässt sich Tabelle C.16 entnehmen, dass 14 Leistungsmaße über beide Aufgaben hinweg eindeutig keine signifikante Abhängigkeit von Systemleistung und Erwartungshaltung zeigen. Dies gilt in vier Fällen auch über die beiden Stichproben SP_A und SP_B hinweg. Mit 11 Variablen besonders stark vertreten sind hier verschiedene Zeitmaße wie die durchschnittliche Betrachtungsdauer eines Dokuments (Z01). Dies lässt darauf schließen, dass innerhalb des hier verwendeten experimentellen Designs die betrachteten Zeitmaße keine hilfreichen Indikatoren für Systemleistung und Erwartungshaltung darstellen.

Tab. 6.10.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System 1, System 2 und Erwartung auf die Benutzerleistung für $A2$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen C.19 und C.24 in Anhang C.3 entnommen werden.

		System 1		System 2		Erwartung	
ID	**Beschreibung**	**$S1_G$**	**$S1_S$**	**$S2_G$**	**$S2_S$**	**E_H**	**E_N**
M03	Anz. aufg. Dok. (erste Suche)	13,42	11,67	**15,84**	**9,25**	11,36	13,73
M05	Anz. aufg. irrel. Dok.	4,58	6,43	**4,10**[a]	**6,90**	4,52	6,48
M07	Anz. falsch irrel. bew. Dok.	2,88	2,40	**3,13**	**2,15**[a]	2,13	3,15
M08	Anz. falsch rel. bew. Dok.	1,88	1,85	**1,20**[a]	**2,53**	1,65	2,08
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	3,38	3,08	**4,10**[a]	**2,35**	3,48	2,98
M12	Anz. rel. bew. Dok. (erste Suche)	8,43	7,83	**10,58**[a]	**5,68**	7,57	8,68
M15	Anz. richtig irrel. bew. Dok.	3,52	4,06	**2,56**	**5,02**[a]	3,54	4,04
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	3,04	2,48	**3,88**[a]	**1,65**	2,98	2,54
M18	Anz. richtig rel. bew. Dok. (erste Suche)	6,63	6,04	**8,46**[a]	**4,21**	6,06	6,61
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	0,28	0,31	**0,18**[a]	**0,42**	0,27	0,32
V02	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	0,72	0,68	**0,80**[a]	**0,59**	0,71	0,68
V06	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	0,51	0,38	**0,57**	**0,33**[a]	0,42	0,48
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	0,11	0,10	**0,06**[a]	**0,14**	0,11	0,10
V09	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	0,16	0,18	**0,10**[a]	**0,24**	0,15	0,19
V10	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. richtig rel. bew. Dok.}}$	0,23	0,29	**0,11**[a]	**0,41**	0,24	0,29
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	0,19	0,20	**0,13**	**0,26**[a]	0,17	0,21
V16	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. falsch irrel. bew. Dok.}}$	1,25	1,70	**0,86**	**2,10**[a]	1,57	1,39
V17	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	0,47	0,53	**0,41**	**0,60**[a]	0,49	0,51
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	0,53	0,52	**0,60**[a]	**0,45**	**0,56**[a]	**0,49**
V31/BP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	0,83	0,84	**0,90**[a]	**0,77**	0,85	0,82
S01	Anz. Suchen	2,08	2,45	**1,78**[a]	**2,75**	2,15	2,38
S05	Zeit zum ersten richtig rel. bew. Dok.	124,29	116,35	**94,77**[a]	**145,88**	130,37	110,27
S05-log	Zeit zum ersten richtig rel. bew. Dok.	4,59	4,51	**4,34**[a]	**4,76**	4,58	4,52

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Auffällig ist, dass insbesondere für die recallorientierten Leistungsmaße (V03, V04, V22, V23, V26, V27, V32/BR u. V33) ein Einfluss der Systemleistung in SP_A für Aufgabe 1 ausgeschlossen werden kann. In den Fällen bei denen eine Analyse möglich ist, trifft dies auch in SP_B zu (V03 u. V04). Diese Beobachtung bestätigt die in Abschnitt 3.2.1 beschriebenen Ergebnisse, die zeigen, dass Nutzer schlechte Systemleistungen in Bezug auf den Recall durch ihr Verhalten kompensieren. Diese Befunde heben den Schwellenwert für einen sichtbaren Einfluss der Systemgüte auf 35 % relativen Systemunterschied für die AvP bzw. 50 % für BPref an. Dies ergänzt

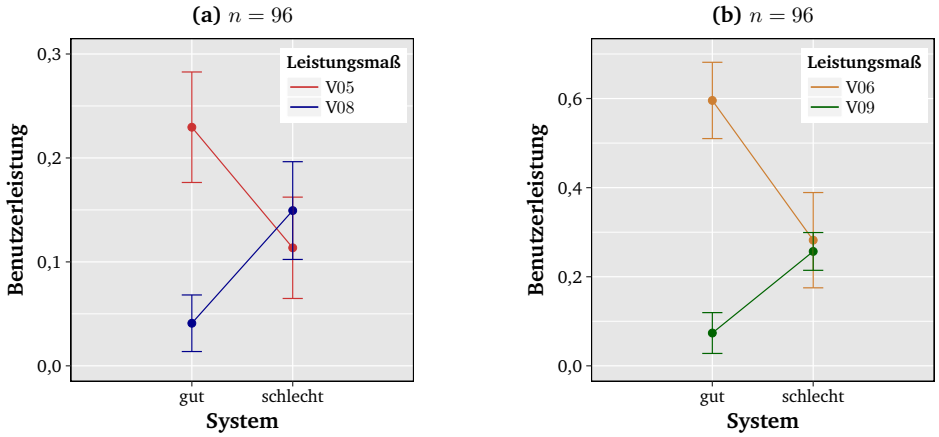


Abb. 6.13.: Systembedingte Anpassung der Relevanzwahrnehmung im zweiten Experiment. Dargestellt sind die Anteile der falsch irrelevant und falsch relevant bewerteten Dokumente in Bezug auf die aufgerufenen Dokumente (a) sowie in Bezug auf die relevant bzw. irrelevant bewerteten Dokumente (b). Im Vergleich zum schlechten System ergibt sich für das gute System jeweils eine Zunahme der falsch irrelevant bewerteten Dokumente (V05 u. V06) bzw. eine Abnahme der falsch relevant bewerteten Dokumente (V08 u. V09). Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

die entsprechenden Ergebnisse von Al-Maskari et al. (2008b) bzw. Al-Maskari et al. (2008a) und Allan et al. (2005) (vgl. Abschn. 6.3.1). Im Einklang mit den Ergebnissen des ersten Experiments kann somit die erste Forschungshypothese (*Bei recallorientierten Leistungsmaßen können Benutzer Unterschiede in der Systemgüte kompensieren.*) erneut bestätigt werden.

Signifikante Effekte der Systemleistung ergeben sich bei der ersten Aufgabe für drei der fünf in Abschnitt 6.3.2 eingeführten Variablengruppen (vgl. Tab. 6.9). Um den Einfluss des Systems auf die Benutzerleistung einfacher beurteilen zu können, ist die jeweils bessere Benutzerleistung mit einer Fußnote gekennzeichnet. Anhand dieser Markierung lässt sich bereits ablesen, dass die meisten Ergebnisse mit den vermuteten Wirkungszusammenhängen übereinstimmen. In den meisten Fällen ist also die Nutzung des besseren Systems tatsächliche mit einer erhöhten bzw. optimierten Benutzerleistung verknüpft. In Bezug auf die untersuchten Dokumentenmengen zeigt sich so bspw., dass weniger irrelevante Dokumente aufgerufen (M05) oder falsch relevant bewertet (M08) werden. Ein analoges Verhalten lässt sich für die meisten Precisionmaße beobachten (V02, V21, V19, V25, V28/PCP u. V31/BP). Besonders ausgeprägt zeigt sich der Systemleistungseffekt auf das Verhältnis aus aufgerufenen relevanten und aufgerufenen Dokumenten (V02), da hier das Ergebnis von SP_A durch SP_B gestützt wird. Dies überrascht nicht, da im Fall der höheren Systemleistung tatsächlich mehr relevante Dokumente in den Ergebnislisten enthalten sind.

Von besonderem Interesse sind Ergebnisse, die den vermuteten Wirkungszusammenhängen widersprechen. Zu nennen sind hier die Variablen M15, B06, V05, V06, V14 und V17, die den bereits im ersten Experiment beobachteten systemleistungsinduzierten Anpassungseffekt bestätigen (vgl. Abschn. 5.4.2). Während V14 und V17 als weitere Precisionmaße anzusehen sind, die erfassen, wie sich das Verhältnis richtig irrelevant bewerteter Dokumente (M15) zu anderen Dokumentenmengen darstellt, sind V05 und V06 Maße für die Ungenauigkeit der Recherche und werden daher als *Imprecisionmaße* bezeichnet (vgl. Abschn. 6.3.2). In allen vier Leistungsmaßen erreichen Benutzer des besseren Systems schlechtere Benutzerleistungswerte. Dabei zeigt sich,

dass Probanden aus der Versuchsgruppe mit dem besseren System mehr Dokumente fälschlich als irrelevant bewerten (V05 u. V06). Die Anteil der Dokumente, die im Gegensatz zu den Juroren als irrelevant bewertet werden, nimmt also mit steigender Systemqualität zu. Der Anteil der richtig als irrelevant bewerteten Dokumente zeigt hingegen entsprechend ein umgekehrtes Verhalten (V14 u. V17). Wie in Abbildung 6.13 veranschaulicht, weisen diese Unterschiede in den Leistungsmaßen V05 und V06 also wiederum auf eine restriktivere bzw. weniger strenge Relevanzbewertung durch Benutzer des besseren bzw. schlechteren Systems hin.

Auch die im Zuge der Nutzung des besseren Systems im Mittel etwas negativer ausfallende Bewertung relevanter Dokumente im Rahmen der letzten durchgeführten Suche (B06) kann in diesem Sinne interpretiert werden. Da relevant bewerteten Dokumenten ein Skalenwert von 1 und irrelevant bewerteten Dokumenten ein Skalenwert von 0 zugeordnet wird, bedeutet ein niedrigerer Wert, dass Benutzer des besseren Systems eine restriktivere Bewertungsstrategie verfolgen und geringere Anzahl relevante Dokumente als relevant bewerten, als dies im umgekehrten Fall auf Benutzer des schlechteren Systems zutrifft. Dabei kann dieser Wert aufgrund der hier verwendeten binären Bewertungsskala auch als Anteil der richtig als relevant bewerteten Dokumente an den aufgerufenen relevanten Dokumenten interpretiert werden. Zusätzlich interessant ist dabei der Aspekt, dass es sich um die letzte durchgeführte Suche handelt, lässt dies doch vermuten, dass die Anpassung der Bewertungsstrategie dynamisch erfolgt.

Auf Basis der bis hierhin diskutierten Einzelergebnisse kann also auch die zweite Forschungshypothese angenommen werden: Bei precisionorientierten Leistungsmaßen passen Benutzer ihre Relevanzdefinition der Systemgüte an. Im Vergleich führt dabei eine höhere Systemleistung zu strengeren bzw. eine niedrigere Systemleistung zu weniger strikten Relevanzkriterien.

Als nächstes werden die Ergebnisse in Bezug auf die zweite Aufgabe betrachtet (vgl. Tab. 6.10). Zunächst fällt auf, dass die Systemgüte der zuerst verwendeten Suchmaschine ($S1$) für keine der Benutzerleistungsvariablen einen signifikanten Einfluss zeigt. Für die Qualität des zur Bearbeitung der zweiten Aufgabe verwendeten Systems ($S2$) lassen sich hingegen signifikante Effekte nachweisen. Für Benutzerleistungsmaße, die in beiden Aufgaben einen signifikanten Effekt der Systemleistung zeigen, bestätigen sich die Resultate der ersten Aufgabe. Aus diesem Grund wird an dieser Stelle der Schwerpunkt auf die Beschreibung der hinzugekommenen Effekte gelegt (M03, M07, M12, M11, M18, M17, V16, S01, S05, S05-log). Wie schon bei der ersten Aufgabe entsprechen auch hier die meisten Ergebnisse den Erwartungen. So werden im Zuge der Nutzung des besseren Systems bspw. mehr Dokumente richtig als relevant bewertet (M18 u. M17). Auch werden im Fall des besseren Systems weniger Suchanfragen benötigt (S01). Unabhängig davon, dass die Anzahl der Suchanfragen natürlich mit der Suchstrategie der Testperson zusammenhängt, deutet dies darauf hin, dass es den Probanden bei Nutzung des besseren Systems leichter fällt, relevante Dokumente zu finden. Dies kann als Hinweis darauf interpretiert werden, dass der beobachtete Kompensationseffekt in Bezug auf recallorientierte Leistungsmaße mit einem erhöhten Aufwand für die Nutzer des schlechteren Systems einhergeht. Diese These wird auch dadurch unterstützt, dass es beim schlechteren System signifikant länger dauert, bis das erste richtig relevant bewertete Dokument gefunden wird (S05). Wie bei der ersten Aufgabe erreichen Benutzer des besseren Systems außerdem bessere Leistungswerte bei V31/BP und V28/PCP. Zusätzlich ergibt sich für die V28/PCP in vier von fünf Stichproben ein Haupteffekt in Bezug auf die

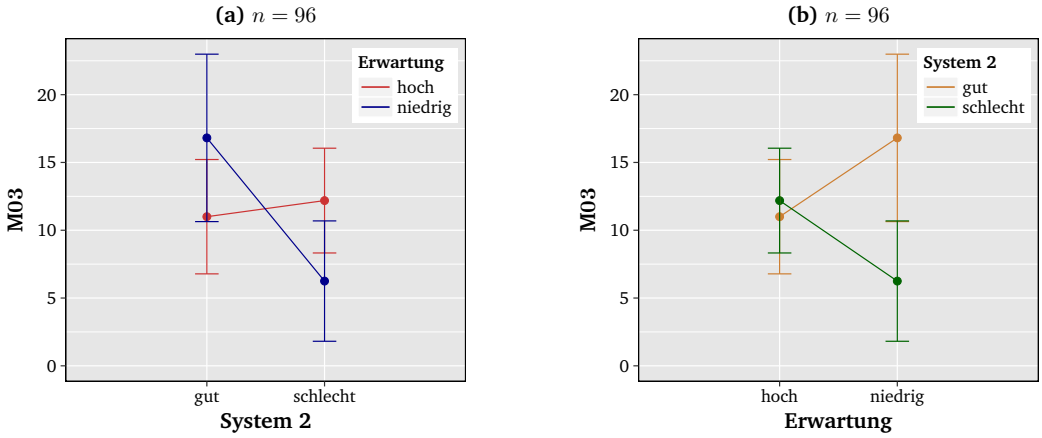


Abb. 6.14.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für die Anzahl der aufgerufenen Dokumente bei der ersten Suche (M03) in $A2$. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte des in Aufgabe 2 verwendeten Suchsystems, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bei Testpersonen mit hoher Erwartungshaltung ist der Einfluss der Systemleistung unterdrückt. Im Fall einer niedrigen Erwartungshaltung hingegen führt die bessere Systemleistung zu einer größeren Anzahl aufgerufener Dokumente im Vergleich zum schlechteren System. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

Erwartungshaltung. Dieser besteht darin, dass Testpersonen, die eine positive Erwartungsmanipulation erhalten, eine bessere Leistung erbringen als Probanden in der Vergleichsgruppe. Dies ist auf eine gegenläufige Abhängigkeit der beiden für sich nicht signifikanten Dokumentenmengen M01 und M16 zurückzuführen aus deren Verhältnis sich die V28/PCP ergibt: Zum einen rufen Probanden mit einer höheren Erwartung weniger Dokumente (M01) und insbesondere weniger relevante Dokumente (M06) auf. Gleichzeitig aber bewerten sie mehr Dokumente in Übereinstimmung mit den Juroren als relevant (M16). Somit lässt die V28/PCP darauf schließen, dass Probanden mit einer hohen Erwartung relevante Dokumente eher als relevant akzeptieren und gleichzeitig weniger Dokumente aufrufen bevor sie davon ausgehen, ihr Informationsbedürfnis gestillt zu haben.

Darüber hinaus wird mit M03 ein Leistungsmaß signifikant, dessen Abhängigkeit von der Systemleistung sich im Vorfeld nicht direkt erschließt. Hier zeigt sich für den Haupteffekt S_2 ein signifikant höherer Wert für das bessere System. Im Fall des besseren Systems werden also mehr Dokumente der ersten Ergebnisliste aufgerufen. Da darüber hinaus jedoch auch die Wechselwirkung zwischen Erwartungshaltung und zweitem System signifikant ist (vgl. Anh. C.3, Tab. C.19 u. Tab. C.24), kann der Haupteffekt nicht allein interpretiert werden. Wie die Wechselwirkungsgraphen in Abbildung 6.14 verdeutlichen, hängt der Einfluss der Systemleistung entscheidend von der Erwartungshaltung der Probanden ab. Bei einer hohen Erwartungshaltung ergibt sich so gut wie kein Unterschied zwischen den beiden Systemen, wohingegen die Probanden bei einer niedrigen Erwartungshaltung im Falle des besseren Systems fast doppelt so viele Dokumente aufrufen, wie im Fall des schlechteren Systems. Dieses Verhalten kann im Sinne des C/D-Paradigmas als ein Übertreffen der Erwartung gedeutet werden, was dazu führt, dass die Probanden viele Dokumente aufrufen, die ihnen als möglicherweise relevant erscheinen.

Wie im Kontext der ersten Aufgabe weisen die signifikanten Mittelwertunterschiede in Bezug auf die Systemleistung für die Leistungsmaße M15, V06, V14 und V17 auf den bereits diskutierten systembedingten Anpassungseffekt der Relevanzwahrnehmung hin. Dieser Befund spiegelt sich nun auch in den Resultaten der Variablen M07 und V16 wider. Nutzer des besseren Systems bewerten mehr Dokumente fälschlicherweise als irrelevant (M07). Dies führt weiterhin dazu, dass das Verhältnis zwischen richtig zu falsch irrelevant bewerteten Dokumenten (V16) für das schlechtere System höher ausfällt. Damit kann der Anpassungseffekt des Relevanzempfindens der Probanden auch in der zweiten Aufgabe bestätigt werden.

Zusammenfassend zeigen die Ergebnisse, dass die beiden in Bezug auf die Benutzerleistung getroffenen Forschungshypothesen bestätigt werden können. Die Systemleistung wirkt sich in der Regel positiv auf die Benutzerleistung aus. Dies gilt bei der Stichprobe SP_A für die erste und zweite Aufgabe gleichermaßen. Der gegenteilige Befund bei bestimmten Precisionmaßen bestätigt die Annahme, dass eine höhere Suchmaschinenqualität zu strengeren Bewertungsmaßstäben bei der Relevanzbewertung führt. Recallmaße hingegen werden von der Systemleistung nicht beeinflusst. Die Vollständigkeit der Suche betreffend sind Benutzer also, wie schon angenommen, in der Lage, trotz geringerer Systemleistung ihr Informationsbedürfnis zu befriedigen. In Bezug auf die Erwartungshaltung lässt sich festhalten, dass sie die Benutzerleistung beeinflussen kann. Generell ist dieser Effekt aber eher gering und tritt nur vereinzelt auf. Im Lichte der Komplexität des Untersuchungsdesigns scheinen somit weitere Experimente erforderlich, um die Befunde dieser Untersuchung besser einschätzen zu können. Des Weiteren kann aufgrund der Datenlage eine Überprüfung der Ergebnisse für die Stichprobe SP_A mit Hilfe der besser kontrollierten Datenmenge SP_B nur für eine geringe Zahl von Leistungsmaßen erfolgen, in diesen Fällen zeigt sich aber eine gute Übereinstimmung, wobei die Überprüfung im Rahmen signifikanter Effekte jedoch auf ein einzelnes Leistungsmaß (V02) beschränkt bleibt.

6.4.4. Skalenbildung

Im Vorfeld der Auswertung der Benutzerzufriedenheit wird eine explorative Faktorenanalyse über alle Zufriedenheitsvariablen durchgeführt mit dem Ziel eine erste Faktorstruktur in den Items zu identifizieren. Diese Verdichtung der Items auf Teilskalen soll zu einer Erleichterung der Auswertung und der Ergebnisinterpretation beitragen.

Als Datengrundlage zur Skalenbildung wird der unbalancierte Datensatz SP_A ohne fehlende Werte ($n = 240$) verwendet, wobei unbalanciert bedeutet, dass pro Versuchsgruppe alle verfügbaren Fälle berücksichtigt werden. Da jeder Teilnehmer zwei Suchaufgaben bearbeitet, gehen 118 Teilnehmer doppelt in die Auswertung ein. In vier Fällen kann aufgrund fehlender Werte nur eine der beiden Aufgaben einbezogen werden. Ein derartiges Vorgehen bringt zwar grundsätzlich die Möglichkeit von systematischen Verzerrungen, wie sie etwa durch die Kumulation von Ausreißerdaten entstehen können, mit sich, jedoch ist andererseits eine nach Aufgaben getrennte Skalenbildung auf Basis der beiden zu testenden Systeme als unbefriedigend anzusehen, da so die Vergleichbarkeit der Teilskalen nicht gegeben wäre. Zudem dient die Faktorenanalyse, wie sie in dieser Arbeit zur Anwendung kommt, vorrangig der Datenreduktion bzw. der Bewältigung der verfügbaren Datenmengen im Vorfeld der Auswertung, im Gegensatz zur Konstruktion und Validierung von Fragebögen.

Historisch bedingt weisen die beiden verwendeten Fragebogenteile (EUCS-Instrument und

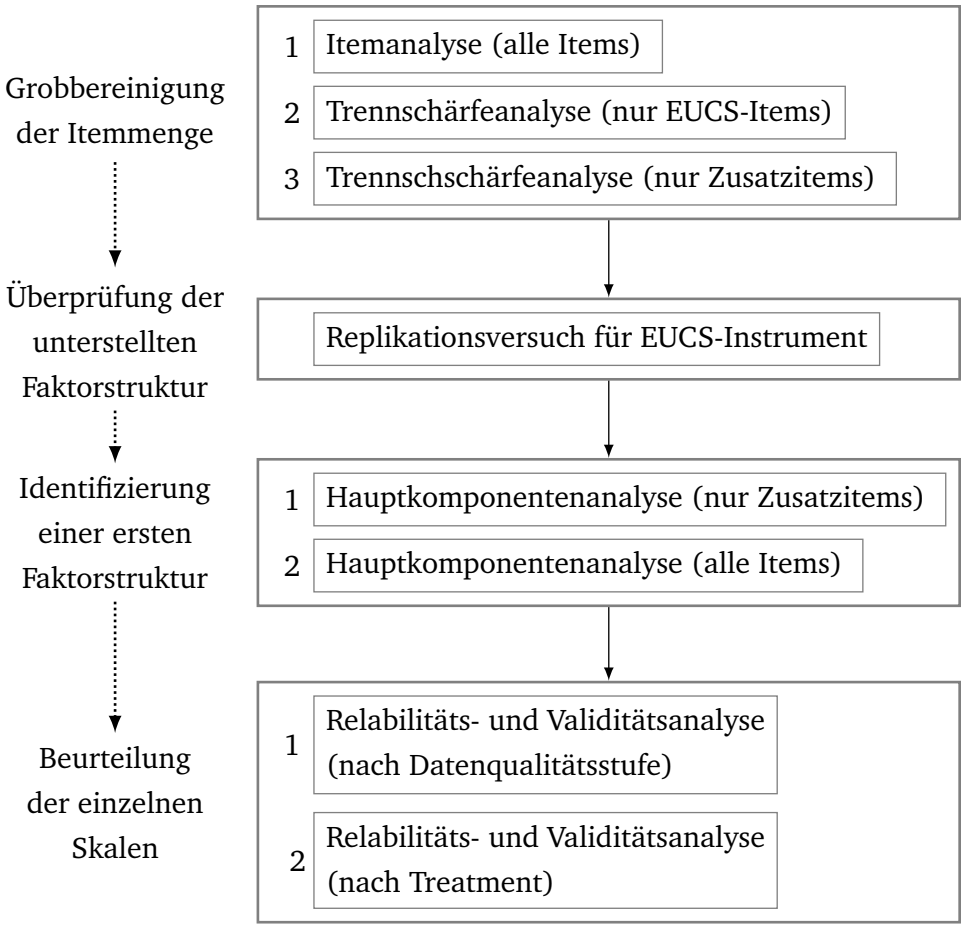


Abb. 6.15.: Vorgehensweise bei der Skalenbildung.

selbst entwickelte Zusatzitems) eine unterschiedliche Skalenbreite auf. Bei den EUCS-Items wird die Beurteilung entsprechend der Vorgabe durch Doll und Torkzadeh (1988) auf einer Skala von 1 bis 5 vorgenommen, wobei die 5 für den höchsten Grad der Zustimmung steht. Bei den Zusatzitems kann die Zustimmung in Anlehnung an das erste Experiment auf einer Skala von 1 bis 7 ausgedrückt werden (vgl. Abschn. 6.3.2). Auch hier wird eine höhere Zustimmung mit einem höheren Skalenwert ausgedrückt. Für unmittelbare Vergleiche der Bewertungen werden diese Skalen im Vorfeld der Skalenbildung mit Hilfe der sog. POMP-Transformation in eine prozentuale Skala überführt (Cohen et al., 1999).

Abbildung 6.15 gibt einen schematischen Überblick über die für die Skalenbildung gewählte Vorgehensweise. In der ersten Phase (vgl. Abschn. 6.4.4.1) wird zunächst eine Itemanalyse über alle verwendeten Zufriedenheitsitems durchgeführt mit dem Ziel, ungeeignete Frageitems aus der weiteren Auswertung auszuschließen. Selektionskriterien sind dabei ein Trennschärfekoeffizient $< 0{,}5$ und eine Kriteriumsvalidiät $< 0{,}4$. Mit Hilfe zweier weiterer Trennschärfeanalysen werden daraufhin die beiden verwendeten Fragebogenteile separat betrachtet. Anschließend erfolgt in der zweiten Phase (vgl. Abschn. 6.4.4.2) ein Replikationsversuch der von Doll und Torkzadeh (1988) beschriebenen Faktoren, um zu überprüfen, ob sich die Struktur des EUCS-Instruments bestätigen lässt. Zu diesem Zweck wird eine Hauptkomponentenanalyse mit Faktorenanzahl und Rotationstechnik identisch zu Doll und Torkzadeh (ebd.) durchgeführt. Zur weiteren Auswertung werden sowohl Mittelwerte (Summe der Items/Anzahl der Items) als auch Faktorwerte der

Skalen berechnet, sodass pro Skala ein gewichteter und ein ungewichteter Skalenwert zur Verfügung steht. In der dritten Phase (vgl. Abschn. 6.4.4.2) erfolgen Hauptkomponentenanalysen sowohl der nicht im EUCS-Instrument enthaltenen Items separat als auch aller Zufriedenheitsitems gemeinsam. Ziel ist es in beiden Fällen, einen ersten Eindruck der zugrunde liegenden Faktorstruktur der Frageitems zu gewinnen. Wie im Fall des Replikationsversuchs wird pro Skala sowohl der Mittelwert als auch der Faktorwert ermittelt. Abschließend werden in einer vierten Phase (vgl. Abschn. 6.4.4.3) die einzelnen Faktoren sukzessive einer Beurteilung unterzogen. Zunächst wird jede Skala nach Datenqualitätsstufe getrennt auf seine Reliabilität und Validität hin untersucht. Als Maß der Reliabilität wird die interne Konsistenz (Cronbachs Alpha) ermittelt. Die Validität der Skalen wird erneut anhand des in Abschnitt 6.4.4.1 beschriebenen Außenkriteriums überprüft.

Als Ergebnis stehen am Ende der Skalenbildung Subskalen aus drei Itemmengen zur Verfügung, die es ermöglichen, systematisch verschiedene, für den Umgang mit Suchmaschinen wichtige Inhaltsbereiche der Benutzerzufriedenheit zu erfassen. Im Folgenden werden zunächst die Ergebnisse der Itemanalyse beschrieben.

6.4.4.1. Itemanalyse

Das Ziel einer Itemanalyse besteht darin, die Eignung der verwendeten Frageitems anhand gegebener Antwortdaten in Bezug auf das zu erfassende Konstrukt, im vorliegenden Fall die Nutzerzufriedenheit, zu evaluieren. Dazu werden der Studie von Doll und Torkzadeh (1988) folgend, zwei Kennzahlen verglichen: die Korrelation zwischen dem Einzelitem und dem Gesamtwert aller Items (Trennschärfe) sowie die Korrelationen jedes Einzelitems mit einem Außenkriterium, von dem angenommen wird, dass es das Konstrukt Benutzerzufriedenheit valide erfasst (Kriteriumsvalidität). Der sog. Trennschärfekoeffizient misst dabei, inwieweit die Antworten auf ein spezielles Frageitem mit dem Gesamtwert des Fragebogens korreliert. Je größer diese Trennschärfe, desto eher ist das Item in der Lage, zwischen einer hohen und einer niedrigen Merkmalsausprägung zu differenzieren. Um eine Überschätzung der Trennschärfe zu vermeiden, wird vor der Berechnung eine Korrektur durchgeführt, bei der der jeweilige Itemwert vom Gesamtwert abgezogen wird. Als Außenkriterium hingegen dienen zwei zusätzliche Fragen nach der allgemeinen Zufriedenheit der Testpersonen mit der Suchmaschine, die ebenfalls von Doll und Torkzadeh (ebd.) übernommen werden können. Der genaue Wortlaut dieser beiden Items lautet: *Ist die Suchmaschine erfolgreich?* (F12), *Sind Sie mit der Suchmaschine zufrieden?* (F13). Die Korrelationen werden in beiden Fällen nach Pearson ermittelt. Um einen Eindruck davon zu bekommen, welche Items für die Skalenbildung weniger geeignet sind, werden die von Doll und Torkzadeh (ebd., S. 264) vorgeschlagenen Schwellenwerte verwendet. Danach sollten Items von der Analyse ausgeschlossen werden, die einen Trennschärfekoeffizient kleiner als $0{,}5$ aufweisen oder deren Korrelation mit dem Außenkriterium geringer als $0{,}4$ ausfällt.

In Tabelle 6.11 werden die Ergebnisse der Itemanalyse aller Zufriedenheitsitems zusammengefasst. Ähnlich der beiden Kriteriumsitems F12 und F13 werden auch die Items F22, F25 und F26 als allgemeine Items zur Erfassung der Gesamtzufriedenheit aufgefasst und bei der Skalenbildung nicht berücksichtigt. Die Ergebnisse bezüglich der Trennschärfe der separat durchgeführten Analysen von EUCS- und Zusatzitems weichen nur geringfügig von den Werten der Gesamtauswertung ab und können daher Anhang C.1 entnommen werden. Wie Tabelle 6.11 zu entnehmen

ist, weisen die verwendeten Zufriedenheitsitems bis auf zwei Ausnahmen insgesamt eine gute Qualität auf und erfassen inhaltlich ähnliche Informationen. Eine Ausnahme bilden die Items F09 und F21, welche mindestens in einem der beiden Fälle unter dem jeweiligen Schwellenwert liegen. F09 stammt aus dem EUCS-Instrument und ist dort der Skala zur Erfassung von Benutzerfreundlichkeit zugeordnet. Bei F21 handelt es sich um eines der Zusatzitems, die entwickelt wurden, um auch stärker kontextbezogene Zufriedenheitsaspekte einbeziehen zu können.

Tab. 6.11.: Trennschärfe und Kriteriumsvalidität aller Zufriedenheitsitems ($n = 240$).

Item	Beschreibung	Korrigierte Item-Total-Korrelation	Korrelation mit dem Kriterium
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	0,77	0,77
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	0,76	0,73
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	0,71	0,69
F04	Liefert die Suchmaschine genügend Information?	0,73	0,72
F05	Ist die Suchmaschine präzise?	0,78	0,74
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	0,84	0,80
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,57	0,54
F08	Ist die Suchmaschine benutzerfreundlich?	0,59	0,58
F09	Ist die Suchmaschine einfach zu bedienen?	**0,42**	0,43
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	0,74	0,70
F11	Liefert die Suchmaschine aktuelle Information?	0,53	0,47
F14	Es war einfach, die Aufgabe zu bearbeiten.	0,64	0,51
F15	Es war einfach, zu dem Thema zu suchen.	0,66	0,56
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	0,84	0,81
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	0,65	0,58
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	0,79	0,74
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	0,81	0,73
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,67	0,53
F21	Ich hatte genügend Zeit, um eine effektive Suche durchzuführen.	**0,46**	**0,29**
F23	Ich bin mit meiner Suchleistung zufrieden.	0,66	0,50
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	0,63	0,58

Als Konsequenz aus dieser Analyse werden diese beiden Items von der Hauptkomponentenanalyse über alle Items ausgeschlossen, nicht jedoch für den Replikationsversuch sowie die Analyse der Zusatzitems. Für die EUCS-Items könnte ein solcher Ausschluss bedeuten, dass die Faktorstruktur des EUCS-Instruments nicht repliziert werden kann. Für die Analyse der Zusatzitems liegt die Trennschärfe für Item 21 mit 0,53 knapp über dem empfohlen Schwellenwert von 0,5 (vgl. Anh. C.1, Tab. C.2), sodass in diesem Fall nur die Korrelation mit dem Außenkriterium mit 0,29 unterhalb des festgelegten Schwellenwertes von 0,4 liegt (vgl. Tab. 6.11). Daher wird dieses Item zunächst in der Itemmenge für die Analyse der Zusatzitems belassen.

6.4.4.2. Explorative Faktorenanalyse

Wie in Abschnitt 6.4.4 beschrieben, erfolgt im ersten Schritt eine Überprüfung der grundsätzlichen Replikationsfähigkeit des EUCS-Instruments. Dazu wird mit den Daten der EUCS-Items eine Hauptkomponentenanalyse (PCA) mit orthogonaler Rotation (varimax) von 5 Faktoren durchgeführt (vgl. Abschn. 4.3.1). Der Kaiser-Meyer-Olkin-Koeffizient (KMO) bestätigt die Eignung der Stichprobe für die Durchführgung einer Faktorenanalyse. Unter Einbeziehung aller 11 Items ergibt sich ein KMO-Wert von 0,91, was auf eine ausgezeichneten Eignung und eine hinreichende Stichprobengröße hinweist (Field et al., 2012, S. 776). Alle KMO-Werte der einzelnen Items sind zudem größer als 0,84, was deutlich über der akzeptablen Grenze von 0,5 liegt. Auch der Bartlett-Test auf Sphärizität der Daten ist signifikant ($\chi^2(55) = 1633$, $p < 0{,}001$), was eine ausreichend hohe Korrelation zwischen den Items anzeigt. Eine weitere Voraussetzung, die

zu überprüfen ist, besteht darin, dass zwischen den Items keine zu hohen Multikollinearitäten bestehen (vgl. Abschn. 4.3.1). Dabei liegt eine Multikollinearität vor, wenn zwischen zwei oder mehr Frageitems so starke Korrelationen bestehen, dass der individuelle Beitrag der Einzelitems zu einem Faktor schwierig zu bestimmen ist. Eine einfache Heuristik, um zu überprüfen, ob eine Multikollinearität vorliegt, setzt voraus, dass die Determinante der Korrelationsmatrix keinen Wert nahe Null ($< 0{,}000\,01$) annimmt (Field et al., 2012, S. 771). Für die Daten der EUCS-Items besitzt die Determinante der Korrelationsmatrix einen Wert von $0{,}001$ und liegt somit über dieser kritischen Grenze. Die vorab an das Datenmaterial gestellten Kriterien zur Berechnung einer Faktorenanalyse sind damit erfüllt.

Im Verlauf der Auswertung lässt sich die Faktorenstruktur des EUCS-Instruments nicht vollständig replizieren. Um eine in sich konsistente Faktorlösung zu erhalten, werden drei Items aus der weiteren Auswertung ausgeschlossen. Als erstes werden F10 und F11 aus der Itemmenge entfernt. Beide Items bilden im Rahmen des EUCS-Instruments gemeinsam die Skala Aktualität. Dieser Ausschluss erscheint sinnvoll, weil die Betrachtung der bis dahin erzielten Faktorlösungen die Vermutung nahe legt, dass das Wort *rechtzeitig* von einigen Probanden überlesen wird und F10 somit eine völlig neue Bedeutung bekommt. Da eine Skala mindestens zwei Items enthalten sollte und F11 nunmehr das einzige Item ist, das die Aktualität der Suchergebnisse adressiert, wird auch diese Frage von der weiteren Analyse ausgeschlossen. In einem letzten Analyseschritt wird schließlich auch das schon während der Itemanalyse aufgrund seiner unzureichenden Trennschärfe auffällig gewordene Item F09 aus der Itemmenge entfernt (vgl. Abschn. 6.4.4.1). Die Reliabilitätsanalyse der Skala, welcher dieses Item zunächst zugeordnet wird, ergibt ein Cronbachs Alpha von $0{,}74$. Dieser Wert erhöht sich auf $0{,}77$, wenn F09 nicht berücksichtigt wird. Da sich die Reliabilität der Skala bei Verzicht auf dieses Item erhöht, wird also auch dieses Item aus der weiteren Auswertung ausgeschlossen. Für eine detailliertere Beschreibung des gesamten Vorgehens sei auf Anhang C.2.1 verwiesen.

Für die hier berichtete finale Lösung mit den verbleibenden 8 EUCS-Items wird nunmehr eine Hauptkomponentenanalyse mit obliquer Rotationsmethode (oblimin) durchgeführt (KMO-Kriterium: $0{,}90$; Bartlett-Test: $p < 0{,}001$) (vgl. Abschn. 6.4.4.1). Eine oblique Rotation scheint vor allem aus zwei Gründen gerechtfertigt. Zum einen können psychologische Konstrukte nur selten als unkorreliert vorausgesetzt werden, zum anderen legen in den Voranalysen auftretende Doppel- und Mehrfachladungen einzelner Items eine Korreliertheit der Faktoren nahe (vgl. Anh. C.2.1, Tab. C.3 - C.5). Tabelle 6.12 zeigt die Faktorladungen der finalen Lösung nach der Rotation. Anhand dieser rotierten Ladungsmatrix lassen sich die drei Subskalen des EUCS-Instruments *Genauigkeit*, *Inhalt* und *Benutzerfreundlichkeit* identifizieren, die zusammen $82{,}09\,\%$ der Gesamtvarianz aufklären. Die internen Konsistenzen (Cronbachs Alpha) der Skalen liegen zwischen $0{,}77$ und $0{,}91$, was insgesamt auf einen akzeptablen bis exzellenten Zusammenhang der Items hinweist. Wenngleich die Zuordnung der einzelnen Items zu den Skalen teilweise von der ursprünglichen Faktorstruktur abweicht, macht das dargestellte Ladungsmuster doch deutlich, dass sich die ursprüngliche Faktorstruktur auch in den vorliegenden Untersuchungsdaten wiederfindet. Abgesehen von Item F01, welches im EUCS-Instrument der Skala Inhalt zugeordnet ist, stimmen die ersten beiden Faktoren (SK01 u. SK02) mit dem Original überein. An die Stelle von Item F09 tritt in Skala SK03 Item F07, welches ursprünglich Bestandteil der Skala Darstellung

ist. Betrachtet man das dargestellte Ladungsmuster im Einzelnen, zeigt sich darüber hinaus, dass Item F03 eine laut Field et al. (ebd., S. 794) als hoch zu bewertende Nebenladung auf der neuen Genauigkeitsskala aufweist. Ihr Auftreten bringt zum Ausdruck, dass die entsprechende Item-Formulierung in der Wahrnehmung der Testteilnehmer inhaltliche Konnotationen beinhaltet, die eine eindeutige Zuordnung zu nur einer der drei extrahierten Komponenten nicht möglich erscheinen lässt. Da die Nebenladung jedoch in diesem Fall inhaltlich nachvollziehbar ist, wird von einem weiteren Ausschluss dieses Items abgesehen. An dieser Stelle sei jedoch darauf verwiesen, dass für die Berechnung der Mittelwerte jedes Item nur einmal verwendet wird und zwar bei dem Faktor, für den es die höchste Faktorladung aufweist (vgl. Abschn. 6.4.4.1). Da pro Skala sowohl der Mittelwert als auch der gewichtete Faktorwert ermittelt wird, besteht im Rahmen der Auswertung somit die Möglichkeit, den Einfluss dieser Skalen mit und ohne den Beitrag von Nebenladungen zu untersuchen. Mögliche Gründe für die gefundenen Abweichungen gegenüber der ursprünglichen Faktorstruktur könnten in der Übersetzung des Fragebogens, aber auch in dem besonderen Kontext des kontrollierten Experiments gesehen werden, in dessen Rahmen alle Probanden das jeweils zu bewertende System zuvor nicht kennen und im Test zum ersten Mal verwenden. Darüber hinaus ist zu beachten, dass das Item *Is the information clear?* des EUCS-Instruments aus übersetzungstechnischen Gründen nicht mit erhoben wird (vgl. Abschn. 6.3.2). Auch dieses Auslassen eines Items kann zu den Verschiebungen in der Faktorstruktur beigetragen haben.

Tab. 6.12.: Ergebnisse der Hauptkomponentenanalyse der EUCS-Items ($n = 240$). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Genauigkeit	Inhalt	Benutzerfreundlichkeit
F05	Ist die Suchmaschine präzise?	**0,97**	−0,08	0,07
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,71**	0,17	0,17
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,62**	0,37	−0,03
F04	Liefert die Suchmaschine genügend Information?	−0,10	**0,92**	0,13
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	0,18	**0,75**	0,03
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,40**	**0,61**	−0,13
F08	Ist die Suchmaschine benutzerfreundlich?	−0,01	0,07	**0,88**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,09	−0,03	**0,86**
	Skalenbezeichnung	SK01	SK02	SK03
	Anteil an Gesamtvarianz (in %)	30,96	29,65	21,48
	α	0,91	0,86	0,77

In Phase 3 des Skalenbildungsprozesses werden die zusätzlich entwickelten Zufriedenheitsitems in die Analyse einbezogen. Dazu werden zunächst die nicht im EUCS-Instrument enthaltenen Items alleine und anschließend alle Zufriedenheitsitems gemeinsam betrachtet (vgl. Abschn. 6.4.4). Tabelle 6.13 fasst die Voraussetzungen zur Eignung beider Datensätze für eine explorative Faktorenanalyse zusammen. Die Werte des KMO-Kriteriums sowie des Bartlett-Tests zeigen, dass die Anforderungen an eine explorative Faktorenanalyse in beiden Fällen erfüllt sind. Die Determinante der Korrelationsmatrix deutet für den Fall, dass alle Items in die Analyse einbezogen werden, an, dass Multikollinearität ein Problem darzustellen scheint. Nachdem zunächst die beiden Items F10 und F11 aus dem Datensatz entfernt werden, die sich bereits in der zuvor beschriebenen Analyse als problematisch erwiesen haben, ergibt eine Analyse der Multikollinearität verursachenden Items, dass sich das Multikollinearitätsproblem durch das zusätzliche Entfernen

von Item F06 beheben lässt.

Tab. 6.13.: Eignung der Daten für eine explorative Faktorenanalyse.

Datensatz	**#Items**	**Kaiser-Meyer-Olkin-Koeffizient** $> \mathbf{0{,}5}$	**Bartlett-Test auf Sphärizität** $\leq \mathbf{0{,}05}$	**Determinante der Korrelationsmatrix** $> \mathbf{1 \cdot 10^{-5}}$
nur Zusatzitems	10	0,898 großartig (great)	$\chi^2(45) = 1397$ $p < 0{,}001$	0,003
alle Items	19	0,950 ausgezeichnet (superb)	$\chi^2(171) = 3346$ $p < 0{,}001$	$5 \cdot 10^{-7}$

Da im Vorfeld keine Annahmen über die zugrunde liegenden Faktoren bestehen, werden in beiden Fällen verschiedene Verfahren (Luhmann, 2013, S. 290 ff.) zur Bestimmung der optimalen Faktorenanzahl herangezogen und vergleichend ausgewertet (vgl. Abschn. 6.4.4.1). Bei der Auswahl der Faktoren bezüglich der Analyse der 10 Zusatzitems bieten sich zwei Alternativen an. Das Eigenwertkriterium nach Jolliffe deutet auf drei Faktoren hin. Joliffe berichtet, dass das Eigenwertkriterium nach Kaiser zu strikt sei und schlägt stattdessen vor alle Faktoren mit einem Eigenwert größer 0,7 beizubehalten (Field et al., 2012, S. 762). Ein ergänzender Scree-Plot bestätigt graphisch die dreifaktorielle Struktur (vgl. Anh. C.2.2, Abb. C.1). Der Very Simple Structure-Wert (VSS) hingegen legt eine Faktorlösung mit vier Faktoren nahe, da Mehrfachladungen auf maximal zwei Faktoren pro Item beschränkt bleiben (Komplexitätsgrad 2). Die übrigen Verfahren zur Bestimmung der Faktorenanzahl (Parallelanalyse, Minimal Average Partial-Kriterium) legen die Extraktion von nur einem Faktor nahe. In Anbetracht der Tatsache, dass alle Frageitems darauf angelegt sind, Benutzerzufriedenheit im Allgemeinen abzufragen, ist es wenig verwunderlich, dass einige Kriterien eine Einfaktorlösung favorisieren. Dennoch erscheint dieser Ansatz wenig hilfreich, da an dieser Stelle das Ziel in der Identifikation spezifischer Komponenten der Benutzerzufriedenheit liegt. Aus diesem Grund werden im Folgenden zwei Hauptkomponentenanalysen mit Oblimin-Rotation und drei oder vier Faktoren durchgeführt und verglichen. Im Endergebnis zeigt sich, dass die Extraktion von vier Faktoren zu der inhaltlich am Besten interpretierbaren Lösung führt. Tabelle 6.14 zeigt die rotierte Ladungsmatrix dieser Lösung, die alternative Dreifaktorenlösung ist in Anhang C.2.2 in Tabelle C.9 aufgeführt.

Die vier Komponenten der favorisierten Lösung klären zusammen 79,99 % der Varianz auf und lassen sich inhaltlich wie folgt beschreiben: Die eindeutig dem Prozess der Suche zuzuordnende erste Komponente setzt sich aus einer Beurteilung der Relevanz der erhaltenen Suchergebnisse und einer Beurteilung der Fähigkeit des Systems, den Benutzer bei der Auswahl relevanter Dokumente angemessen zu unterstützen, zusammen. Die zweite Komponente umfasst Items, die ausschließlich die Schwierigkeit der zu bearbeitenden Aufgabe beschreiben. Demgegenüber beschreibt die dritte Komponente die Selbstwahrnehmung der Probanden während der Aufgabenbearbeitung. Hier sind die Beurteilung der verfügbaren Bearbeitungszeit und die Beurteilung der eigenen Suchleistung als Variablen enthalten. In der vierten und letzten identifizierten Komponente schließlich dominiert die Beurteilung des Systems in Bezug auf Nutzungserlebnis und Nachvollziehbarkeit. Die internen Konsistenzen der einzelnen Skalen können, mit Ausnahme der

Tab. 6.14.: Ergebnisse der Hauptkomponentenanalyse der Zusatzitems ($n = 240$). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Suche	Aufgabe	Eigenleistung	Benutzerfreundlichkeit
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,91**	0,08	0,03	−0,08
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,81**	−0,07	0,10	0,17
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	**0,63**	0,23	0,03	0,17
F14	Es war einfach, die Aufgabe zu bearbeiten.	−0,04	**0,90**	0,11	−0,03
F15	Es war einfach, zu dem Thema zu suchen.	0,04	**0,90**	−0,09	0,06
F21	Ich hatte genügend Zeit, um eine effektive Suche durchzuführen.	−0,07	−0,01	**0,89**	0,10
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,30	0,07	**0,62**	0,02
F23	Ich bin mit meiner Suchleistung zufrieden.	0,29	0,18	**0,61**	−0,06
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	−0,07	0,06	0,16	**0,85**
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	0,35	0,04	−0,15	**0,66**
	Skalenbezeichnung	SK04	SK05	SK06	SK07
	Anteil an Gesamtvarianz (in %)	26,32	19,94	18,79	14,93
	α	0,90	0,81	0,80	0,67

zuletzt genannten Skala (α =0,67), als gut bis exzellent beurteilt werden ($\alpha > 0{,}8$).

In Bezug auf die Hauptkomponentenanalyse aller Zufriedenheitsitems spricht das Eigenwertkriterium ($> 0{,}7$) für eine vierstufige Faktorlösung, der VSS-Wert Komplexitätsgrad 2 für drei Faktoren. Die übrigen Extraktionskriterien legen hingegen erneut eine einfaktorielle Struktur nahe. Zieht man weiterhin die Ergebnisse aus den ersten beiden Analysen zurate, so wäre auch eine fünffaktorielle Faktorlösung mit den Faktoren Inhalt, Genauigkeit, Benutzerfreundlichkeit, Aufgabe und Eigenleistung ein denkbares Resultat. Im Folgenden werden daher drei Hauptkomponentenanalysen mit drei, vier und fünf Faktoren durchgeführt. Da die inhaltliche Interpretation sowohl bei der Drei- wie auch bei der Fünffaktorenlösung nicht für alle Einzelfaktoren zufriedenstellend ist (vgl. Anh. C.2.3), fällt die Entscheidung aufgrund der eindeutigeren und damit besser interpretierbaren Struktur auf die vierfaktorielle Lösung. Mit diesen vier Faktoren werden insgesamt $73{,}57\,\%$ der Varianz aufgeklärt. Tabelle 6.15 zeigt die Ladungsmatrix der Vierkomponentenlösung für die 16 Zufriedenheitsitems nach der Oblimin-Rotation (KMO-Kriterium: $0{,}94$; Bartlett-Test: $p < 0{,}001$).

Die vier Komponenten lassen sich wie folgt interpretieren: Auf der ersten Komponente laden diejenigen Items des EUCS-Instruments hoch, die im Rahmen des Replikationsversuchs die ersten beiden Faktoren (Inhalt u. Genauigkeit) ausmachen (vgl. Tab. 6.12). Zusätzlich laden vier Items aus der Menge der Zusatzitems auf die erste Komponente, welche ebenfalls im weiteren Sinne die Qualität der Suchergebnisse zum Thema haben. Man könnte diese Komponente also als Sucherfolg und Zufriedenheit mit der Suche bezeichnen. Auf der zweiten Komponente laden Items, die sich auf die Bedienung und das Design der Suchmaschine beziehen. Diese Komponente lässt sich erneut unter dem Begriff Benutzerfreundlichkeit zusammenfassen. Die dritte Komponente entspricht der im Rahmen der Analyse der Zusatzitems als Zufriedenheit mit der gestellten Aufgabe interpretierten Skala (vgl. Tab. 6.14). Auf der letzten Komponente schließlich laden erneut Items, welche die Wahrnehmung der eigenen Suchleistung beschreiben. Die internen Konsistenzen der einzelnen Skalen sind als akzeptabel bis exzellent einzuschätzen.

Zusammenfassend lässt sich sagen, dass die theoretische Faktorenstruktur des EUCS-Instru-

Tab. 6.15.: Ergebnisse der Hauptkomponentenanalyse aller Items ($n = 240$). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Suche	Benutzer-freundlich-keit	Aufgabe	Eigen-leistung
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,93**	−0,13	−0,08	0,05
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,89**	−0,01	0,10	−0,11
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,80**	−0,01	−0,05	0,15
F05	Ist die Suchmaschine präzise?	**0,77**	0,13	−0,01	0,00
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,73**	−0,02	0,14	0,08
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,69**	0,17	0,01	0,08
F04	Liefert die Suchmaschine genügend Information?	**0,68**	0,05	0,06	0,08
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	**0,55**	0,21	0,24	0,06
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	**0,53**	0,35	0,03	−0,10
F08	Ist die Suchmaschine benutzerfreundlich?	0,02	**0,88**	0,07	−0,09
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,04	**0,81**	−0,06	0,07
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	0,00	**0,68**	0,02	0,26
F15	Es war einfach, zu dem Thema zu suchen.	0,14	0,00	**0,87**	−0,08
F14	Es war einfach, die Aufgabe zu bearbeiten.	−0,07	0,02	**0,86**	0,18
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,12	0,08	0,00	**0,79**
F23	Ich bin mit meiner Suchleistung zufrieden.	0,07	0,00	0,15	**0,77**
	Skalenbezeichnung	SK08	SK09	SK10	SK11
	Anteil an Gesamtvarianz (in %)	35,01	15,39	12,19	10,98
	α	0,94	0,79	0,81	0,76

ments in der explorativen Faktorenanalyse in Teilen repliziert werden kann. Dies spricht für die Validität des EUCS-Instruments auch im Kontext des IR und zeigt, dass es eine gute Entscheidung darstellt, dieses bewährte Fragebogen-Instrument zugrunde zu legen. Durch die zusätzlich entwickelten Items können Anhaltspunkte für weitere Dimensionen der Benutzerzufriedenheit gewonnen und einer ersten Prüfung unterzogen werden. Die faktoranalytisch ermittelten Skalen lassen sich inhaltlich gut interpretieren und sind ausreichend reliabel. Über die im Rahmen der Faktoranalyse entwickelten Skalen hinaus, werden auch gemittelte Zufriedenheitsurteile getrennt nach EUCS- (SK-E-09) und Zusatzitems (SK-Z-09) sowie in Bezug auf alle verwendeten Zufriedenheitsitems zusammen (SK-G-09) gebildet. Eine Übersicht über alle in Experiment 2 verwendeten Skalen kann Tabelle B.8 in Anhang B.8 entnommen werden. Die im Folgenden dargestellte Analyse befasst sich mit der Beurteilung der einzelnen Skalen im Hinblick auf die in Abschnitt 6.4.2 beschriebenen Teilstichproben.

6.4.4.3. Reliabilitäts- und Validitätsanalyse

Neben der bereits berichteten Reliabilität der Skalen bezogen auf alle 240 in die explorative Faktorenanalyse einbezogenen Fälle, soll an dieser Stelle noch einmal genauer auf die in dieser Stichprobe vertretenen Teilstichproben eingegangen werden. Mit diesem mehrperspektivischen Ansatz ist die in Abschnitt 6.4.2 diskutierte Frage nach der Gültigkeit der in diesem Kapitel berichteten Untersuchungsergebnisse verbunden. Stimmen die Werte für die verschiedenen hier betrachteten Fallgruppen überein, kann dies sowohl als zusätzlicher Hinweis für die Gültigkeit der verwendeten Skalen als auch als Referenz für die Eignung der Gesamtstichprobe SP_A interpretiert werden.

Die Tabellen 6.16 und 6.17 beschreiben die interne Konsistenz sowie die kriteriumsbezogene Validität der einzelnen Skalen nach Aufgabe und Datenqualitätsstufe getrennt. Die Gütekriterien

der Skalen sind allgemein als gut zu beurteilen: Die Reliabilitätskoeffizienten für die betrachteten vier Fallgruppen liegen im Mittel bei etwa $0{,}82$. Ausnahmen bilden die Skalen SK03 und SK07 für die dieser Koeffizient in vier Fällen knapp unter dem geforderten Mindestwert von $0{,}7$ liegt. Die Validität betreffend gibt es in allen betrachteten Fallgruppen erwartungskonforme Zusammenhänge zu dem herangezogenen Außenkriterium. Alle Korrelationen liegen hier über dem von Doll und Torkzadeh (1988, S. 264) vorgeschlagenen Schwellenwert von $0{,}4$ (vgl. Abschn. 6.4.4.1). Im Vergleich zwischen der bereinigten (SP_B) und der nicht bereinigten Stichprobe (SP_A) bleiben sowohl die ermittelten Reliabilitätskoeffizienten als auch die Validitätskoeffizienten weitgehend stabil. Die größte Differenz ergibt sich im Kontext der zweiten Aufgabe. Hier beträgt die Differenz zwischen SP_B und SP_A für die Kriteriumsvalidität von Skala SK11 $0{,}14$, was jedoch weiterhin als unauffällig zu bewerten ist. Auch im Vergleich zwischen erster und zweiter Aufgabe liegt die höchste Differenz bei $0{,}14$. Konkret handelt es sich dabei erneut um Skala SK03, für die in der Stichprobe SP_B beide Gütekriterien bei der zweiten Aufgabe etwas schlechter ausfallen. Jedoch sind auch diese Unterschiede marginal, sodass die Zuverlässigkeit der ermittelten Skalen insgesamt gesehen als durchaus zufriedenstellend eingeschätzt werden kann.

Tab. 6.16.: Skalenreliabilität und Kriteriumsvalidität nach Datenqualität für $A1$.

Skala	**Beschreibung**	**#Items**	**Cronbachs Alpha**		**Kriteriumsvalidität**	
			SP_A	**SP_B**	**SP_A**	**SP_B**
			$n = 119$	$n = 54$	$n = 119$	$n = 54$
SK01	Genauigkeit	3	0,90	0,90	0,84	0,80
SK02	Inhalt	3	0,85	0,85	0,77	0,78
SK03	Benutzerfreundlichkeit	2	0,81	0,77	0,64	0,75
SK04	Suche	3	0,87	0,88	0,85	0,83
SK05/10	Aufgabe	2	0,76	0,77	0,58	0,60
SK06	Eigenleistung	3	0,81	0,85	0,52	0,46
SK07	Benutzerfreundlichkeit	2	0,68	0,70	0,65	0,61
SK08	Suche	9	0,93	0,93	0,88	0,85
SK09	Benutzerfreundlichkeit	3	0,82	0,81	0,68	0,79
SK11	Eigenleistung	2	0,76	0,83	0,57	0,51

Tab. 6.17.: Skalenreliabilität und Kriteriumsvalidität nach Datenqualität für $A2$.

Skala	**Beschreibung**	**#Items**	**Cronbachs Alpha**		**Kriteriumsvalidität**	
			SP_A	**SP_B**	**SP_A**	**SP_B**
			$n = 121$	$n = 55$	$n = 121$	$n = 55$
SK01	Genauigkeit	3	0,91	0,91	0,84	0,84
SK02	Inhalt	3	0,87	0,83	0,86	0,77
SK03	Benutzerfreundlichkeit	2	0,71	0,63	0,62	0,61
SK04	Suche	3	0,93	0,91	0,81	0,80
SK05/10	Aufgabe	2	0,86	0,88	0,59	0,49
SK06	Eigenleistung	3	0,75	0,77	0,55	0,44
SK07	Benutzerfreundlichkeit	2	0,66	0,62	0,71	0,74
SK08	Suche	9	0,95	0,94	0,88	0,85
SK09	Benutzerfreundlichkeit	3	0,75	0,77	0,68	0,67
SK11	Eigenleistung	2	0,75	0,79	0,61	0,47

Darüber hinaus wird in Bezug auf den Einfluss der Datenqualität untersucht, ob der Ausschluss bestimmter kritischer Fallgruppen die Skalenreliabilität und die Kriteriumsvalidität beeinflussen und inwiefern die hier diskutierten Gütekriterien auch hinsichtlich der Originalskalen des EUCS-Instruments erfüllt sind. Auch die hier gewonnenen Ergebnisse sprechen für die Eignung der

Gesamtstichprobe und sind im Detail in Anhang C.2.4 dokumentiert.

Zusammenfassend lässt sich festhalten, dass die in diesem Abschnitt diskutierten Gütekriterien weitestgehend erfüllt sind. Zudem wird die Eignung der Gesamtstichprobe SP_A durch die Tatsache unterstrichen, dass sich die berichteten Gütekriterien nicht übermäßig sensitiv gegenüber der Einbeziehung unterschiedlicher Fallgruppen erweisen.

6.4.5. Auswertung der Benutzerzufriedenheit

Der folgende Abschnitt stellt die Ergebnisse des zweiten Experiments in Bezug auf die Benutzerzufriedenheit dar. In die Betrachtung einbezogen werden dabei sowohl die 26 in Abschnitt 6.3.2 beschriebenen Einzelitems als auch die im Rahmen des zweiten Experiments erarbeiteten Zufriedenheitsskalen (vgl. Abschn. 6.4.4). Die Auswertung erfolgt analog zur Benutzerleistung nach Aufgaben getrennt (vgl. Abschn. 6.4.3). Die entsprechenden Ergebnisse sind in den Tabellen 6.18 und 6.19 zusammengefasst. Auch in diesem Fall werden ausschließlich eindeutig oder in der Tendenz signifikante Ergebnisse (fett hervorgehoben) berichtet, wobei für jede Variable die Stichprobe mit dem signifikantesten Ergebnis ausgewählt wird. Weiterführende Informationen bezüglich der verwendeten Varianzanalyse (klassisch vs. robust) oder der Qualität des Effekts (eindeutig vs. tendenziell) sowie die entsprechenden Signifikanzniveaus sind in den Tabellen C.23 und C.24 in Anhang C.3 dargestellt. Wie schon bei der Darstellung der Benutzerleistung werden im Folgenden die Befunde anhand ausgewählter Zufriedenheitsmaße erläutert. Weitere, die jeweilige Interpretation stützende, Items sind hingegen zur besseren Dokumentation der Ergebnisse in Klammern angegeben. Dabei gilt erneut, dass diese Auflistungen als Maß für die Stabilität der Effekte interpretiert werden sollten, die zum Verständnis der Befunde jedoch nicht im einzelnen nachvollzogen werden müssen.

Wie schon bei der Benutzerleistung stellt sich erneut das Problem, dass wegen des Auftretens von Topiceffekten, die eine zusätzliche Topicbalancierung erforderlich machen würden, die Stichprobe SP_B nur für sechs Variablen ausreichend groß ist, um sinnvoll eine Varianzanalyse durchführen zu können (vgl. Tab. 6.8). Allerdings zeigen diese Variablen, vermutlich aufgrund der immer noch geringen Stichprobengröße, in drei Fällen eindeutig keine signifikanten Effekte (F04, F17 u. F18) und in drei Fällen uneindeutige Befunde (F05, SK07-M u. SK-T). Aus diesem Grund ist die folgende Darstellung auf die Stichprobe SP_A beschränkt. Insgesamt können 51 signifikante Effekte bei 42 unterschiedlichen Frageitems und Zufriedenheitsskalen nachgewiesen werden. Davon entfallen 37 Effekte auf die erste und 14 auf die zweite Aufgabe. Dabei liegt die Stichprobengröße für Aufgabe 1 bei mindestens 96, für Aufgabe 2 bei mindestens 80 Probanden (vgl. Anh. C.3). Demgegenüber stehen drei Variablen (F10, F21 u. SK06-F) bei denen für beide Aufgaben eindeutig keine Abhängigkeit von Systemleistung und Erwartungshaltung nachgewiesen werden kann und weitere vier bzw. acht Variablen, bei denen dies nur auf die erste bzw. zweite Aufgabe zutrifft (vgl. Anh. C.3, Tab. C.16).

Der interessanteste Unterschied zwischen den Ergebnissen der beiden Aufgaben zeigt sich in der Natur der auftretenden Effekte. Während bei der ersten Aufgabe ausschließlich der Einfluss der Erwartungshaltung und nie die Systemgüte eindeutig signifikant wird, kehrt sich dieses Verhalten bei der zweiten Aufgabe vollständig um. In diesem Fall wird nur noch die Systemleistung signifikant, wohingegen kein nachweisbarer Einfluss der Erwartungshaltung zu finden ist. Ein messbarer Einfluss des ersten Systems ist somit bei keiner der Aufgaben zu finden. Es

Tab. 6.18.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerzufriedenheit für $A1$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen C.20 und C.23 in Anhang C.3 entnommen werden.

		System		Erwartung	
ID	**Beschreibung**	**S_G**	**S_S**	**E_H**	**E_N**
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	3,67	3,96	**4,13**[c]	**3,50**
F05	Ist die Suchmaschine präzise?	3,47	3,11	**3,63**[c]	**2,95**
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	3,54	3,33	**3,85**[c]	**3,02**
F08	Ist die Suchmaschine benutzerfreundlich?	3,63	3,78	**4,21**[c]	**3,19**
F09	Ist die Suchmaschine einfach zu bedienen?	4,50	4,54	**4,73**[c]	**4,31**
F11	Liefert die Suchmaschine aktuelle Information?	3,52	3,87	**4,02**[c]	**3,37**
F12	Ist die Suchmaschine erfolgreich?	3,77	3,92	**4,13**[c]	**3,56**
F13	Sind Sie mit der Suchmaschine zufrieden?	3,80	3,57	**4,09**[c]	**3,28**
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	4,73	4,79	**5,21**[c]	**4,31**
F22	Ich bin mit den Suchergebnissen zufrieden.	4,64	4,67	**5,11**[c]	**4,21**
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	4,71	4,44	**5,15**[c]	**4,00**
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	4,17	4,05	**4,84**[c]	**3,38**
SK01-M	Genauigkeit	0,63	0,53	**0,66**[c]	**0,50**
SK01-F	Genauigkeit	0,09	-0,005	**0,36**[c]	**-0,28**
SK02-M	Inhalt	0,62	0,64	**0,70**[c]	**0,56**
SK02-F	Inhalt	-0,12	-0,02	**0,23**[c]	**-0,37**
SK03-M	Benutzerfreundlichkeit	0,61	0,57	**0,71**[c]	**0,48**
SK03-F	Benutzerfreundlichkeit	0,02	-0,02	**0,47**[c]	**-0,47**
SK04-M	Suche	0,65	0,69	**0,73**[c]	**0,62**
SK04-F	Suche	0,02	0,13	**0,35**[c]	**-0,19**
SK06-M	Eigenleistung	0,47	0,50	**0,53**[c]	**0,43**
SK07-M[a]	Benutzerfreundlichkeit	0,62	0,61	**0,68**[c]	**0,55**
SK07-F	Benutzerfreundlichkeit	-0,05	-0,23	**0,19**[c]	**-0,47**
SK08-M	Suche	0,65	0,62	**0,71**[c]	**0,56**
SK08-F	Suche	-0,05	0,09	**0,35**[c]	**-0,31**
SK09-M	Benutzerfreundlichkeit	0,62	0,58	**0,70**[c]	**0,50**
SK09-F	Benutzerfreundlichkeit	0,02	-0,10	**0,40**[c]	**-0,47**
SK-A	Accuracy (EUCS)	0,60	0,55	**0,69**[c]	**0,45**
SK-C	Content (EUCS)	0,62	0,66	**0,70**[c]	**0,57**
SK-E[b]	Ease of Use (EUCS)	0,76	0,81	**0,88**[c]	**0,69**
SK-T	Timeliness (EUCS)	0,71	0,73	**0,79**[c]	**0,65**
SK-K	Kriteriumsskala	0,70	0,74	**0,81**[c]	**0,64**
SK-E-88	EUCS-Skala-1988	0,67	0,69	**0,75**[c]	**0,61**
SK-E-09	EUCS-Skala-2009	0,64	0,59	**0,70**[c]	**0,53**
SK-Z-09	Zusatzskala-2009	0,61	0,61	**0,66**[c]	**0,56**
SK-G-09	Gesamtskala-2009	0,64	0,63	**0,69**[c]	**0,57**

[a] Entspricht auch der Skala SK18-M.
[b] Entspricht auch der Skala SK13-M.
[c] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit.

lässt sich also feststellen, dass die Erwartungshaltung zu Beginn einen starken Einfluss auf das Zufriedenheitsurteil der Probanden ausübt, welcher jedoch über die Zeit abzunehmen scheint. Demgegenüber überwiegen bei der zweiten Aufgabe die tatsächlich mit dem System gemachten Erfahrungen. Interessant ist in diesem Zusammenhang, dass die Systemgüte des ersten Systems bei der Auswertung der zweiten Aufgabe keinen zusätzlichen Einfluss zeigt, vielmehr scheint das Zufriedenheitsurteil nur die unmittelbar mit dem (zweiten) System gemachten Erfahrungen widerzuspiegeln. Dies deutet auf eine dynamische Abhängigkeit der Benutzerzufriedenheit hin, die im Verlauf dieser Arbeit im Rahmen des dritten Experiments noch genauer untersucht wird (vgl. Kap. 7). Auch stehen diese Ergebnisse im Einklang mit den Beobachtungen von Szajna und Scamell (1993) die feststellen, dass unrealistische Erwartungen von Testpersonen mit der Zeit durch tatsächlich gemachte Erfahrungen verdrängt werden. Als weiteres Indiz für diesen Erklä-

Tab. 6.19.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System 1, System 2 und Erwartung auf die Benutzerzufriedenheit für $A2$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen C.21 und C.24 in Anhang C.3 entnommen werden.

		System 1		System 2		Erwartung	
ID	**Beschreibung**	**$S1_G$**	**$S1_S$**	**$S2_G$**	**$S2_S$**	**E_H**	**E_N**
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	3,63	3,55	**3,97**[c]	**3,20**	3,68	3,50
F12	Ist die Suchmaschine erfolgreich?	3,65	3,85	**4,02**[c]	**3,48**	3,88	3,63
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	4,70	4,95	**5,30**[c]	**4,35**	5,05	4,60
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	4,65	4,70	**5,10**[c]	**4,25**	4,85	4,50
SK04-F	Suche	-0,18	0,04	**0,29**[c]	**-0,42**	0,12	-0,25
SK05-F	Aufgabe	-0,05	-0,02	**0,17**[c]	**-0,25**	-0,18	0,11
SK07-M[a]	Benutzerfreundlichkeit	0,62	0,64	**0,69**[c]	**0,56**	0,60	0,65
SK08-F	Suche	-0,02	-0,13	**0,26**[c]	**-0,41**	0,04	-0,19
SK11-M[b]	Eigenleistung	0,57	0,56	**0,61**[c]	**0,51**	0,57	0,56
SK11-F	Eigenleistung	0,09	0,06	**0,36**[c]	**-0,22**	-0,13	0,28
SK-A	Accuracy (EUCS)	0,56	0,61	**0,67**[c]	**0,50**	0,62	0,55
SK-K	Kriteriumsskala	0,64	0,63	**0,73**[c]	**0,54**	0,69	0,58
SK-Z-09	Zusatzskala-2009	0,63	0,64	**0,68**[c]	**0,59**	0,62	0,65
SK-G-09	Gesamtskala-2009	0,63	0,65	**0,70**[c]	**0,57**	0,64	0,64

[a] Entspricht auch der Skala SK18-M.
[b] Entspricht auch den Skalen SK15-M und SK19-M.
[c] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit.

rungsansatz kann außerdem der fehlende Einfluss der ursprünglichen Erwartungsmanipulation auf die im Anschluss an die zweite Aufgabe im Rahmen des Gedankenexperiments zu beantwortenden Erwartungsitems E02 und E04 (vgl. Tab. C.16) angeführt werden. Allerdings zeigt sich hier auch kein signifikanter Einfluss des zweiten Systems auf die Erwartungsabfrage, was jedoch dem komplexen Untersuchungsdesign mit zwei unterschiedlichen Systemgüten geschuldet sein könnte.

Betrachtet man den tatsächlichen Effekt, den die Erwartungshaltung auf die Benutzerzufriedenheit ausübt, so zeigt sich, dass bei der ersten Aufgabe eine höhere Erwartungshaltung zu einer größeren Zufriedenheit führt. Die Probanden spiegeln in ihren Zufriedenheitsurteilen somit ihre erhaltene Erwartungsmanipulation wider. Dieses Ergebnis stimmt nicht direkt mit den Vorhersagen des C/D-Paradigmas überein, das neben einer Wechselwirkung zwischen Systemleistung und Erwartungshaltung insbesondere eine höhere Zufriedenheit bei der geringeren Erwartungshaltung voraussagt. Dies könnte darauf hindeuten, dass die Erwartungshaltung im Kontext von IR-Systemen eher zu einem Placebo-Effekt bzw. Bestätigungsfehler führt (Habel et al., 2016). Im Sinne der Ergebnisse von Boulding et al. (1993) könnten diese Ergebnisse auch darauf hinweisen, dass die gewählte Form der Erwartungsmanipulation gerade die normativen Erwartungen der Probanden beeinflusst, da diese im Kontext der Kundenzufriedenheit positiv mit dem Zufriedenheitsurteil korrelieren (vgl. Abschn. 3.3.1.4). Im Kontext der signifikanten Systemleistungseffekte der zweiten Aufgabe trägt das bessere System jedoch wie erwartet und in Übereinstimmung mit dem C/D-Paradigma zu einer erhöhten Zufriedenheit bei. Zusammenfassend kann also die dritte Forschungshypothese (*Die Zufriedenheit der Benutzer wird durch ihre Erwartungshaltung und die Systemgüte gemäß den Annahmen des C/D-Paradigmas beeinflusst*) nur in Bezug auf die Systemleistung bestätigt werden.

Durch eine weitergehende Analyse der inhaltlichen Ausrichtung der in den Tabellen 6.18 und 6.19 aufgeführten Items und Skalen lassen sich darüber hinaus folgende Fragen untersuchen: Gibt es reihenfolgespezifische Unterschiede in der Verteilung der Zufriedenheitsvariablen? Welche Aspekte der Zufriedenheit weisen einen stärkeren Zusammenhang mit der Erwartungshaltung auf und welche Aspekte sind stärker an die Systemleistung gekoppelt? Diese Fragen sollen im Folgenden erörtert werden. Reihenfolgespezifische Unterschiede ergeben sich z.B. in der Benutzerfreundlichkeit. Hier werden im Kontext der ersten Aufgabe vier Skalen (SK03-M/SK03-F, SK07-M/SK07-F, SK09-M/SK09-F, SK-E) und zwei Frageitems (F08 u. F09), im Kontext der zweiten Aufgabe hingegen nur eine Skala (SK07-M) signifikant. Die wahrgenommene Benutzerfreundlichkeit des Systems scheint also direkt mit der Erwartungshaltung der Nutzer zusammenzuhängen: Eine höhere Erwartung führt zu einer höheren Zufriedenheit. Geht der Einfluss der Erwartungshaltung zurück, verschwinden somit auch die messbaren Unterschiede bezüglich der Benutzerfreundlichkeit, da sich die beiden Systeme tatsächlich auch nur in der Farbe der Benutzeroberfläche unterscheiden. Dies kann als weiterer Hinweis auf den oben genannten Placebo-Effekt gedeutet werden. Unterstrichen wird diese Beobachtung noch durch die Tatsache, dass die bereits erwähnte Skala zur Benutzerfreundlichkeit SK03-F für die erste Aufgabe eindeutig signifikant (vgl. Tab. 6.18), für die zweite Aufgabe hingegen eindeutig nicht signifikant ist (vgl. Tab. C.16).

Ein ähnliches Verhalten zeigt sich im Kontext der Gesamtzufriedenheit der Testpersonen für die Items F22 und F25, die für die erste Aufgabe einen tendenziellen bzw. eindeutigen Einfluss der Erwartungshaltung aufweisen, jedoch eindeutig nicht signifikant im Rahmen der zweiten Aufgabe werden. Eine gegenläufige Tendenz ist für die wahrgenommene Eigenleistung zu erkennen. Hier zeigt sich eine stärkere Abhängigkeit für die zweite im Vergleich zur ersten Aufgabe. So sind Item F18 und die Skala SK11-F für die erste Aufgabe eindeutig nicht signifikant, während sie für die zweite Aufgabe eine eindeutig bzw. tendenziell signifikante Systemabhängigkeit zeigen.

Ausgehend von dieser Diskussion lassen sich somit vier zentrale Ergebnisse festhalten: 1. *Die Manipulation der Erwartungshaltung ist erfolgreich.* Eine höhere Erwartungshaltung der Testpersonen führt im Rahmen der ersten Aufgabe zu signifikant besseren Zufriedenheitswerten. Das gewählte Testszenario eines fiktiven Vergleichs von zwei unterschiedlichen Systemen scheint demnach gut geeignet zu sein, um die Erwartungshaltung von Testpersonen zu beeinflussen. 2. *Der Einfluss der Erwartungshaltung auf die Benutzerzufriedenheit nimmt mit zunehmender Übung und Erfahrung ab.* Dies deutet auf eine dynamische Abhängigkeit der Benutzerzufriedenheit hin und könnte auch eine Erklärung für den nicht vorhandenen Erwartungseinfluss im ersten Experiment sein, da die Zufriedenheit der Testpersonen dort erst nach Bearbeitung aller und nicht nach jeder einzelnen Aufgabe abgefragt wird. Allerdings ist zu beachten, dass das zweite Experiment in seinem Aufbau recht komplex ist, sodass auch die Tatsache, dass pro Aufgabe eine unterschiedliche Erwartungsmanipulation vorliegt, dazu geführt haben könnte, dass das Treatment der zweiten Aufgabe nicht erinnert wird. Vor dem Hintergrund, dass im ersten Experiment jedoch kein Erwartungseffekt auftritt, erscheint die Interpretation einer dynamischen Erwartungskorrektur aber durchaus plausibel. 3. *Die Erwartungshaltung steht in engem Zusammenhang mit der wahrgenommenen Benutzerfreundlichkeit eines Systems.* Dieses Ergebnis ist aus zwei Gründen bemerkenswert: Zum einen zeigt es deutlich, dass die Erwartung eine wesentliche Einflussgröße

für die Benutzerzufriedenheit darstellt, unterscheiden sich beide Systeme doch nur in der Farbe der Benutzeroberfläche. Dies stützt erneut die Idee eines Placebo-Effekts durch die Erwartungshaltung, bei dem die tatsächliche Systemqualität ausgeblendet wird. Zum anderen machen die Beobachtungen deutlich, dass der Zeitpunkt der Messung Einfluss auf die Messwerte hat. In einem Usability-Test sollte eine Zufriedenheitsmessung also immer erst nach einer längeren Einarbeitungsphase erfolgen, um eine Überlagerung der eigentlichen Zufriedenheitsreaktion durch die Erwartungshaltung der Testpersonen möglichst zu vermeiden. 4. *Die Vorhersagen des C/D-Modells lassen sich im Kontext des zweiten Experiments nicht auf den Prozess der Informationssuche übertragen.* Wie bereits beschrieben, steht die beobachtete Korrelation zwischen Erwartungshaltung und Benutzerzufriedenheit einer Interpretation im Sinne des C/D-Paradigmas entgegen. Vielmehr deuten die Ergebnisse auf das bereits erwähnte Auftreten eines Placebo-Effekts bzw. Bestätigungsfehlers und einer Beeinflussung der normativen Erwartungen durch das Erwartungstreatment hin.

Zusammenfassend lässt sich somit festhalten, dass das Hauptanliegen des zweiten Experiments, eine nachweisbare Manipulation der Erwartungshaltung herbeizuführen, erreicht worden ist und sich durch das beobachtete Abflauen dieses Einflusses eine interessante neue Forschungsperspektive in Bezug auf diese Dynamik eröffnet hat. Im Folgenden werden die Ergebnisse einer Überprüfung der Gütekriterien des Experiments unterzogen und diese zusammenfassend dargestellt.

6.4.6. Überprüfung der Gütekriterien des Experiments

Auch im Zuge der Auswertung des zweiten Experiments wird überprüft, ob untersuchungsbedingte oder personenbezogene Faktoren die Ergebnisse systematisch beeinflussen. Ausgehend von diesen Betrachtungen wird abschließend entschieden, welcher Datensatz für die Auswertung des Experiments am Besten geeignet ist und ob außer den beiden unabhängigen Variablen noch weitere Einflussgrößen als Kovariaten in die eigentliche Auswertung einbezogen werden sollten.

6.4.6.1. Untersuchungsbedingte Störfaktoren

Als erstes untersuchungsbedingtes Gütekriterium wird der Einfluss der beiden Testaufgaben untersucht. Diese Analyse gibt Auskunft darüber, ob die gestellten Aufgaben im Vergleich einfach oder schwer zu lösen sind. Sollte der Vergleich zeigen, dass Probanden eine der Aufgaben signifikant besser lösen oder sie nach ihrer Bearbeitung signifikant zufriedener sind als Teilnehmer, die die andere Aufgabe bearbeitet haben, erfolgt die Balancierung der entsprechenden unabhängigen Variable in den Datensätzen SP_A und SP_B sowohl nach Versuchsgruppe als auch nach Suchthema (SP_{MT}), um den Topiceffekt im Rahmen der Hauptanalyse statistisch zu kontrollieren (vgl. Abschn. 6.4.2).

Um den Einfluss der Suchthemen auf die abhängigen Variablen Benutzerleistung und Benutzerzufriedenheit zu überprüfen, werden einfaktorielle Varianzanalysen mit dem Aufgabenthema als unabhängige Variable durchgeführt. Die Fallauswahl für die vergleichende Analyse erfolgt anhand eines mehrstufigen Verfahrens mit dem Ziel, in jeder Versuchsgruppe die Anzahl der Testpersonen für jede der beiden Aufgaben auszubalancieren. Dazu wird zunächst pro Datensatz (SP_A u. SP_B) und unabhängiger Variable die maximale Gruppengröße bestimmt, für die eine solche Balancierung möglich ist. Dies entspricht gerade der Kombination aus Versuchsgruppe

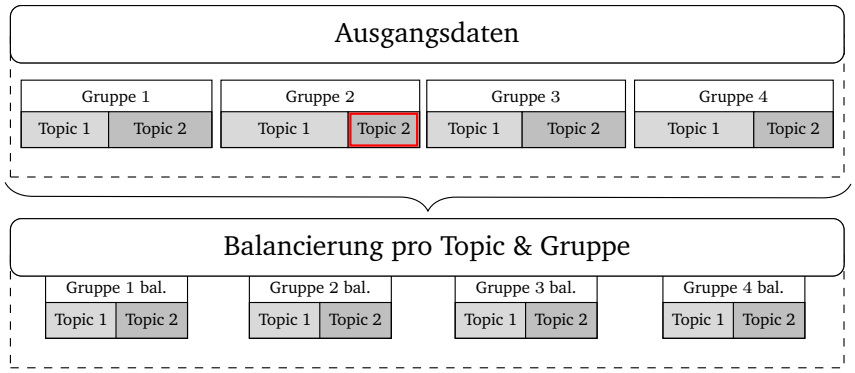


Abb. 6.16.: Vorgehensweise zur Topicbalancierung des Datensatzes nach Versuchsgruppen. Über alle Untersuchungsgruppen hinweg wird zunächst das Topic mit der geringsten Fallzahl ermittelt (rot markiert). Im zweiten Schritt werden in diesem Umfang Zufallsstichproben für jede der Untersuchungsgruppen/Topic-Kombinationen gezogen, was zu einer in Bezug auf Topic und Treatment ausgeglichenen Teilstichprobe führt.

und Topic mit der geringsten Häufigkeit. Diese Anzahl wird daher im nächsten Schritt in der randomisierten Fallauswahl für jede der Topic-Versuchsgruppen-Kombinationen zugrunde gelegt. Um alle möglichen Effekte identifizieren und so eine möglichst konservative Überprüfung des Aufgabeneffekts gewährleisten zu können, wird anders als bei der Fallauswahl für die Hauptauswertung keine Mindestgruppengröße festgelegt, die erreicht werden muss, um die entsprechende Variable in die Auswertung einzubeziehen. In gleicher Weise wie bei der Hauptauswertung erfolgt die Signifikanzprüfung abhängig davon, ob die Voraussetzungen für eine Varianzanalyse erfüllt sind entweder anhand des klassischen oder einer robusten Variante dieses Verfahrens (vgl. Abschn. 6.4.3). Wiederum werden nur diejenigen Effekte berücksichtigt, die in mindestens vier der fünf Stichproben nachweisbar sind. Im Folgenden werden die Ergebnisse knapp dargestellt, detailliertere Informationen zu den Ergebnissen der einzelnen Varianzanalysen können Anhang C.4 entnommen werden.

Tab. 6.20.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A1$ in SP_A. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

							Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M05	Anz. aufg. irrel. Dok.	96	R	1/45	11,38	0,0015	4,96	2,85[a]
V07	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. richtig irrel. bew. Dok.}}$	72	R	1/35	7,28	0,011	0,75[a]	1,38

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.

Signifikante Unterschiede zeigen sich für die Gesamtstichprobe SP_A sowohl bezüglich der Leistungsmaße als auch der Zufriedenheit der Benutzer. Während sich jedoch bei der ersten Testaufgabe nur bei insgesamt sechs Variablen signifikante Unterschiede in den Reaktionen auf das Aufgabenthema feststellen lassen, sind es bei der zweiten Testaufgabe 36 Variablen. Eine mögliche Erklärung könnte sein, dass der Lernaufwand bei der ersten Aufgabe noch überwiegt,

während die Testpersonen bei der zweiten Aufgabe schon eingearbeitet sind und sich mit den Arbeitsabläufen und dem Testsystem besser auskennen. Für die bereinigte Stichprobe SP_B sind die vorliegenden Fallzahlen sehr gering. Sie bewegen sich für die erste Aufgabe zwischen 8 und 16 Teilnehmern, während für die zweite Aufgabe ausnahmslos 8 Probanden pro Suchthema vorhanden sind. Dementsprechend lassen sich für Aufgabe 1 in nur zwei Fällen und für Aufgabe 2 in keinem Fall signifikante Topiceffekte nachweisen. Ob dies der besseren Kontrolle der Stichprobe oder den geringen Fallzahlen geschuldet ist, kann jedoch nicht abschließend geklärt werden. Vor diesem Hintergrund bestätigt sich die konservative Vorgehensweise, die Qualitätsstufe SP_{OT}, bei der ausschließlich nach Versuchsgruppenzugehörigkeit balanciert wird (vgl. Abschn. 6.4.2), nur für solche Variablen zu wählen, bei denen ein Topiceffekt eindeutig ausgeschlossen werden kann. Eine Übersicht der betreffenden Variablen findet sich in Tabelle C.25 im Anhang.

In den Tabellen 6.20 und 6.21 sind die signifikanten Ergebnisse der einfaktoriellen Varianzanalysen der Benutzerleistungsvariablen für Aufgabe 1 bzw. Aufgabe 2 wiedergegeben. Aufgrund des oben beschriebenen Verfahrens zur Fallauswahl bewegen sich die Stichprobenumfänge für die hier betrachteten abhängigen Variablen zwischen 64 und 96 Probanden, sodass mindestens 32 Versuchspersonen pro Suchthema für die Auswertung zur Verfügung stehen. Um die Lesbarkeit der Ergebnisse zu vereinfachen, sind den Tabellen Fußnoten zur Kennzeichnung der jeweils leichteren Aufgabe hinzugefügt. Weiterhin ist zu beachten, dass von den fünf pro Variable untersuchten Stichproben jeweils diejenige mit dem niedrigsten p-Wert berichtet wird. In der Mehrzahl der Fälle sind die Verteilungsvoraussetzungen für eine klassische Varianzanalyse (K) nicht erfüllt, sodass für diese Variablen eine robuste (R) Varianzanalyse durchgeführt wird.

Die Ergebnisse in Bezug Aufgabe 1 machen bereits deutlich, dass keines der beiden Suchthemen eindeutig als leichter bzw. schwerer zu identifizieren ist. So stellt sich das Windthema in Bezug auf die Dokumentenmenge M05 als leichter dar, während für das Verhältnismaß V07 beim Sonnentopic bessere Leistungen erzielt werden. Gleiches gilt auch für die Befunde in Bezug auf Aufgabe 2. So lassen sich für die erste Variablengruppe, die verschiedene Dokumentenmengen umfasst, fünf signifikante Aufgabeneffekte nachweisen. Vier dieser Mittelwertunterschiede (M07, M10, M13 u. M16) legen nahe, dass das Sonnenenergiethema einfacher zu bearbeiten ist, während das Ergebnis in Bezug auf die falsch als relevant bewerteten Dokumenten (M08) den gegenteiligen Effekt andeutet. Der Trend, im Kontext des Sonnenthemas insgesamt mehr Dokumente als relevant zu akzeptieren, setzt sich auch für die anderen Variablengruppen fort. Die Probanden bewerten Dokumente für diese Aufgabe insgesamt positiver (B01, B04, B05, B06), womit der Anteil richtig als relevant bewerteter Dokumente höher (z.B. V29, V32 u. V33) und im Gegenzug der Anteil fälschlicherweise als irrelevant bewerteter Dokumente geringer ausfällt (z.B. V05 u. V07). Gleichzeitig bewerten die Testteilnehmer für das Windenergiethema weniger Dokumente im Widerspruch zu den Juroren als relevant (V08 u. V15). Insgesamt können diese Befunde als aufgabenspezifischer Anpassungseffekt des Relevanzurteils gewertet werden.

Dieser Anpassungseffekt lässt sich auch im Kontext der Zufriedenheitsindikatoren beobachten, deren Ergebnisse in Tabelle 6.22 für Aufgabe 1 und in Tabelle 6.23 für Aufgabe 2 zusammengefasst sind. Hier sind die Teilnehmer durchgehend zufriedener mit dem Sonnenenergiethema, also mit der Suchaufgabe, deren Dokumente sie insgesamt als relevanter bewertetet haben. Die positivere Wahrnehmung der Suchergebnisse ist somit direkt mit einem höheren Zufriedenheits-

Tab. 6.21.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A2$ in SP_A. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

							Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M07	Anz. falsch irrel. bew. Dok.	96	R	1/48	9,95	0,0028	2,25[a]	3,98
M08	Anz. falsch rel. bew. Dok.	96	R	1/44	11,40	0,0015	2,79	1,23[a]
M10	Anz. rel. bew. Dok.	96	R	1/46	10,05	0,0027	11,35[a]	8,15
M13	Anz. rel. bew. Dok. (letzte Suche)	96	R	1/41	8,70	0,0052	10,77[a]	6,29
M16	Anz. richtig rel. bew. Dok.	96	R	1/54	6,67	0,013	9,98[a]	6,92
B01	Durchschn. Bew. irrel. Dok.	64	R	1/36	10,68	0,0024	0,48	0,24[a]
B04	Durchschn. Bew. rel. Dok.	96	R	1/51	21,54	$2{,}5 \cdot 10^{-05}$	0,81[a]	0,64
B05	Durchschn. Bew. rel. Dok. (erste Suche)	80	R	1/44	14,02	0,00052	0,81[a]	0,63
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	80	R	1/44	14,09	0,00051	0,82[a]	0,65
Z11-log	Durchschn. Betrachtungsz.richtig rel. bew. Dok.	96	K	1/94	9,30	0,003	2,79[a]	3,21
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	R	1/44	15,73	0,00026	0,13[a]	0,25
V07	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. richtig irrel. bew. Dok.}}$	64	R	1/24	8,20	0,0084	0,74[a]	1,66
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	R	1/54	12,23	0,00094	0,15	0,076[a]
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	R/K	1/50	21,49	$2{,}5 \cdot 10^{-05}$	0,30[a]	0,47
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	R/K	1/51	27,35	$3{,}2 \cdot 10^{-06}$	0,70[a]	0,53
V13	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$	64	K	1/62	15,45	0,00022	0,76[a]	0,63
V15	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. irrel. Dok.}}$	64	R	1/36	16,28	0,00027	0,62	0,79[a]
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	96	K/R	1/94	16,36	0,00011	0,79[a]	0,65
V32/BR	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	96	R	1/44	7,59	0,0085	0,15[a]	0,11
V33	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	96	R	1/45	9,48	0,0035	0,19[a]	0,13

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.

empfinden verknüpft.

Im nächsten Schritt wird über den Einfluss des Suchthemas hinaus überprüft, ob der Ausschluss bzw. die Einbeziehung problematischer Fallgruppen die Untersuchungsergebnisse maßgeblich beeinflusst. Dazu werden die, bereits in Abschnitt 6.4.1 eingeführten Fallgruppen unterschieden: auffällige Suchbegriffe (SP_{SB}), Manipulation versagt (SP_{MV}) und Test durchschaut (SP_{TD}). Des Weiteren wird untersucht, ob sich die erreichten Leistungs- und Zufriedenheitswerte von Probanden, die nicht die maximale Bearbeitungszeit von zehn Minuten in Anspruch nehmen (SP_{UZ}), systematisch von den Messwerten der übrigen Probanden unterscheiden. Ähnlich wie schon im Zusammenhang mit der Reliabilitäts- und Validitätsanalyse der Skalen geschehen, wird deshalb bei der Analyse möglicher Aufgabeneffekte zusätzlich überprüft, ob der Ausschluss oben genannter Fallgruppen die Ergebnisse der Hauptauswertung in Bezug auf Topiceffekte grundlegend

Tab. 6.22.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A1$ in SP_A. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	96	R	1/37	5,90	0,02	5,69[a]	4,83
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	96	R	1/39	15,65	0,00031	5,31[a]	4,38
SK04-M	Suche	96	R	1/51	10,37	0,0022	0,71[a]	0,60
SK08-M	Suche	96	R	1/55	6,87	0,011	0,70[a]	0,60

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.

verändert (vgl. Abschn. 6.4.4.3). Die Tabellen 6.24 und 6.25 enthalten zu diesem Zweck eine Übersicht, aus der hervorgeht, welche Topiceffekte der Gesamtstichprobe SP_A unter Ausschluss der einzelnen problematischen Fallgruppen reproduziert werden können. Die ausführlichen Ergebnisse sind hingegen in Anhang C.4 dargestellt.

Als erstes fällt auf, dass sowohl für Aufgabe 1 als auch für Aufgabe 2 die Stichprobe SP_B und der Ausschluss der Fallgruppe SP_{SB} zur geringsten Überschneidung mit den in SP_A nachgewiesenen Topiceffekten führt und auch nur wenige zusätzliche Effekte nachweisbar sind. Zunächst deuten diese Befunde darauf hin, dass das Auftreten von Topiceffekten eng mit den von den Probanden verwendeten Suchbegriffen zusammenhängt, da ein Ausschluss von SP_{SB} ja gerade zu ihrem Verschwinden führt. Allerdings erscheint diese Interpretation in Bezug auf die Benutzerleistungsmaße wenig überzeugend, da bei dem verwendeten Suchsystem die Ergebnislisten ja gerade unabhängig von der speziellen Suchanfrage erzeugt werden. Die näher liegende Erklärung ist somit, dass das Verschwinden der Topiceffekte den geringen Stichprobengrößen geschuldet sein könnte. Dies wird gestützt durch die Tatsache, dass sich die Topiceffekte gegenüber dem Ausschluss der übrigen drei Fallgruppen SP_{MV}, SP_{TD} und SP_{UZ}, die jeweils weniger Probanden umfassen, als stabiler herausstellen. Hier werden in allen drei Fällen nahezu fünfzig Prozent der signifikanten abhängigen Variablen reproduziert. Darüber hinaus lässt sich anhand der Tabellen 6.24 und 6.25 ein differenzierterer Eindruck von der Qualität der einzelnen Topiceffekte gewinnen. So weisen die beiden Benutzerleistungsvariablen V05 und V29 in allen vier Auswertungen einen eindeutigen Topiceffekt auf, was für einen sehr stabilen Effekt spricht. Für die abhängigen Variablen V12, B04, M10 und F26 ist dies zumindest tendenziell wahr. Im Gegensatz dazu scheinen die im Rahmen der Hauptauswertung identifizierten Effekte der Variablen V15 und SK-C weniger stabil zu sein, denn in allen vier Zusatzauswertungen finden sich keine signifikanten Abweichungen zwischen den beiden Suchthemen. Ein weiterer interessanter Aspekt ist, dass der Ausschluss der Fallgruppe SP_{UZ} nicht, wie vielleicht zu vermuten wäre, zu einer Reduktion zufriedenheitsbezogener Effekte führt. Vielmehr bleiben auch nach Ausschluss der frühzeitig die Aufgabe beendenden Probanden die meisten der signifikanten Zufriedenheitsunterschiede bestehen. Damit ist zunächst nicht offensichtlich, ob die Zufriedenheit der Versuchspersonen einen direkten Einfluss auf das Abbruchverhalten hat. Für zwei Drittel der Variablen reproduzieren mindestens zwei der drei Fallgrupppen die zuvor identifizierten Topiceffekte. Insgesamt bestätigen diese Befunde, dass ein Balancieren der Treatmentgruppen in Bezug auf die bearbeiteten

Tab. 6.23.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A2$ in SP_A. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	96	R	1/55	8,22	0,0058	3,83[a]	3,19
F08	Ist die Suchmaschine benutzerfreundlich?	96	R	1/57	12,61	0,00077	4,08[a]	3,58
F12	Ist die Suchmaschine erfolgreich?	96	R	1/52	7,78	0,0074	3,98[a]	3,40
F13	Sind Sie mit der Suchmaschine zufrieden?	96	R	1/56	11,64	0,0012	4,00[a]	3,12
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	96	R	1/48	9,07	0,0041	5,42[a]	4,85
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	96	R	1/45	22,40	$2{,}2 \cdot 10^{-05}$	5,19[a]	4,44
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	96	R	1/57	13,23	0,00059	5,08[a]	4,35
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	96	R	1/57	19,83	$4 \cdot 10^{-05}$	5,10[a]	4,06
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	96	R	1/57	15,16	0,00026	4,60[a]	3,42
SK01-M	Genauigkeit	96	R	1/55	9,36	0,0034	0,68[a]	0,52
SK04-M	Suche	96	R	1/40	11,34	0,0017	0,72[a]	0,59
SK08-M	Suche	96	R	1/44	8,66	0,0051	0,68[a]	0,57
SK09-M	Benutzerfreundlichkeit	96	R	1/50	6,55	0,014	0,69[a]	0,57
SK-C	Content (EUCS)	96	R	1/47	6,56	0,014	0,70[a]	0,60
SK-E-09	EUCS-Skala-2009	96	R/K	1/57	10,93	0,0016	0,69[a]	0,57
SK-K	Kriteriumsskala	96	R	1/49	13,06	0,00071	0,72[a]	0,57

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.

Suchthemen in der Hauptauswertung angeraten ist.

Anders als im ersten Experiment erscheint eine Analyse möglicher Reihenfolgeeffekte im Kontext des zweiten Experiments weniger sinnvoll, da die Systemleistung für die beiden Aufgaben variiert wird. Es ist somit schwierig einen möglichen Reihenfolgeeffekt in Bezug auf die Suchaufgabe von dem Treatment in Bezug auf das Suchsystem zu unterscheiden. Zusammenfassend kann somit wie im ersten Experiment festgestellt werden, dass der Schwierigkeitsgrad zwischen den beiden Testaufgaben variiert. Dies kann aber, wie bereits diskutiert (vgl. Abschn. 5.4.4), eher als positive Voraussetzung für die Generalisierbarkeit der Untersuchungsergebnisse gesehen werden, da auch in realen Anwendungssituationen unterschiedlich schwere Aufgaben bewältigt werden müssen. Gleichzeitig zeigt sich, dass diese Schwankungen im Schwierigkeitsgrad nicht nur hinsichtlich der objektiv messbaren Benutzerleistung sondern auch in Bezug auf die subjektiv wahrgenommene Benutzerzufriedenheit beobachtet werden können. Weiterhin lässt sich aus der beschriebenen Stabilität der Ergebnisse in Bezug auf die reduzierten Fallgruppen wie schon im Zusammenhang mit der Reliabilitäts- und Validitätsanalyse eine Eignung der Gesamtstichprobe ableiten. Aufgrund der Gesamtheit dieser Befunde scheint sowohl die Randomisierung der Aufgabenreihenfolge als auch die nach Versuchsgruppe und Suchthema balancierte Fallauswahl aus methodischer Sicht sinnvoll und notwendig gewesen zu sein.

6.4.6.2. Personenbezogene Störfaktoren

Um eventuell auftretenden personenbezogenen Verzerrungen Rechnung zu tragen, werden im Kontext der Gesamtstichprobe Kovarianzanalysen für probandenspezifische Kovariaten durchge-

Tab. 6.24.: Übersicht über beobachtete Topiceffekte für $A1$ in SP_A. Zusätzlich dargestellt sind Topiceffekte, die in SP_B oder unter Ausschluss der Fallgruppen SP_{SB}, SP_{MV}, SP_{TD} und SP_{UZ} eindeutig (E) oder in der Tendenz (T) bestehen bleiben. Der untere Bereich der Tabelle zeigt darüber hinaus die Anzahl der in den betrachteten Teilstichproben zusätzlich hinzukommenden Effekte.

ID	SP_A $n = 72 - 96$	SP_B $n = 24 - 32$	$SP_A \backslash SP_{SB}$ $n = 32 - 40$	$SP_A \backslash SP_{MV}$ $n = 72 - 96$	$SP_A \backslash SP_{TD}$ $n = 64 - 88$	$SP_A \backslash SP_{UZ}$ $n = 56 - 80$
M05	T	-	-	-	T	-
V07	T	-	-	-	-	-
F16	T	-	-	-	-	-
F19	T	-	-	E	T	E
SK04-M	E	-	-	E	-	E
SK08-M	T	-	-	E	-	E
BL	vorhanden	0/2	0/2	0/2	1/2	0/2
	zusätzlich	2	1	4	5	0
	gesamt	2	1	4	6	0
BZ	vorhanden	0/4	0/4	3/4	1/4	3/4
	zusätzlich	0	0	4	0	1
	gesamt	0	0	7	1	4

führt. Dabei werden die folgenden Faktoren in die Betrachtung einbezogen (vgl. Abschn. 6.3.3): das Alter (K01), das Geschlecht (K02), die Muttersprache (K03), ein Set von acht Erfahrungsvariablen (Computernutzungsjahre (K04), Computernutzungsstunden (K05), Domänenwissen (K06), Selbsteinschätzung Domänenwissen (K07), Selbsteinschschätzung Suchmaschinenwissen (K08), Suchmaschinennutzungsjahre (K09), Suchmaschinennutzungsstunden (K10), Suchmaschinenwissen (K11) sowie die Ausgangsmotivation (K12) der Testpersonen. Ausgewählte Ergebnisse der Kovarianzanalysen sind in diesem Abschnitt zusammengefasst. Ausgehend von 146 Variablen und 12 Kovariaten ergeben sich pro Testaufgabe im Höchstfall 35.040 Analysen. Diese hohe Zahl ist der Tatsache geschuldet, dass pro Kombination von Variable und Kovariate sowohl für SP_A als auch für SP_B mindestens fünf unterschiedliche Stichproben zu berücksichtigen sind, für die im Falle von Voraussetzungsverletzungen zusätzlich robuste Tests durchgeführt werden müssen.

Um hier zu einer handhabbaren Anzahl von Analysen zu kommen, beschränkt sich die folgende Darstellung auf die Gesamtstichprobe SP_A. Des Weiteren werden im Folgenden nur Kovarianzanalysen betrachtet, die alle statistischen Voraussetzungen erfüllen und darüber hinaus eine zumindest tendenzielle Aussage über alle fünf Stichproben zulassen. Neben den üblichen Verteilungsvoraussetzungen, deren Verletzung eine robuste Auswertung erforderlich macht, muss im Rahmen einer Kovarianzanalyse die zusätzliche Annahme überprüft werden, dass keine Abhängigkeit zwischen den Kovariaten und Treatments besteht (vgl. Abschn. 4.3.2.3). Ist diese Annahme verletzt, muss die korrespondierende Kombination von unabhängiger Variablen und Kovariate von der weiteren Analyse ausgeschlossen werden. Gleiches gilt, wenn die Voraussetzung homogener Regressionskoeffizienten verletzt ist, was impliziert, dass Interaktionen zwischen den Treatments und der Kovariate vorliegen (vgl. Abschn. 4.3.2.3). Darüber hinaus ist es selbstverständlich erforderlich, dass die Kovariate nicht über alle Testpersonen hinweg konstant ist. Nach Ausschluss all dieser Einschränkungen, bleiben noch 11.560 Tests für die Auswertung der ersten und 4.790 für die Auswertung der zweiten Aufgabe übrig. Da hier sowohl balancierte und unbalancierte Stichproben, als auch klassische und robuste Analysen enthalten sind, reduziert sich

Tab. 6.25.: Übersicht über beobachtete Topiceffekte für $A2$ in SP_A. Zusätzlich dargestellt sind Topiceffekte, die unter Ausschluss der Fallgruppen SP_{SB}, SP_{MV}, SP_{TD} und SP_{UZ} eindeutig (E) oder in der Tendenz (T) bestehen bleiben. Der untere Bereich der Tabelle zeigt darüber hinaus die Anzahl der in den betrachteten Teilstichproben zusätzlich hinzukommenden Effekte.

	$\mathbf{SP_A}$	$\mathbf{SP_A \setminus SP_{SB}}$	$\mathbf{SP_A \setminus SP_{MV}}$	$\mathbf{SP_A \setminus SP_{TD}}$	$\mathbf{SP_A \setminus SP_{UZ}}$
ID	$n = 64 - 96$	$n = 16 - 32$	$n = 64 - 96$	$n = 48 - 80$	$n = 32 - 64$
M07	T	-	T	E	-
M08	E	-	E	-	T
M10	E	T	E	T	E
M13	E	-	T	-	-
M16	E	-	E	-	-
B01	T	-	T	-	-
B04	E	T	E	E	T
B05	E	-	E	T	-
B06	E	-	E	E	-
Z11-log	T	-	E	-	-
V05	E	E	E	E	E
V07	E	-	E	E	-
V08	E	-	E	-	T
V11	E	-	E	E	T
V12	E	T	E	E	T
V13	E	-	T	-	T
V15	E	-	-	-	-
V29	E	E	E	E	E
V32/BR	T	-	E	-	-
V33	T	-	E	-	-
F03	T	-	E	-	E
F08	E	-	T	T	-
F12	E	-	T	-	T
F13	E	T	E	-	E
F16	E	-	E	-	T
F18	E	-	E	E	E
F24	T	-	E	T	T
F25	E	-	T	-	E
F26	E	T	E	E	T
SK01-M	T	-	E	-	E
SK04-M	E	-	T	-	E
SK08-M	E	-	-	T	E
SK09-M	E	T	-	-	-
SK-C	E	-	-	-	T
SK-K	E	-	E	T	E
SK-E-09	T	-	-	-	T
BL	vorhanden	5/20	19/20	10/20	9/20
	zusätzlich	0	4	1	1
	gesamt	5	23	11	10
BZ	vorhanden	3/16	12/16	6/16	14/16
	zusätzlich	1	1	0	3
	gesamt	4	13	6	17

die Anzahl tatsächlich zu betrachtender Analysen schließlich auf insgesamt 4.885 Tests im Fall von Aufgabe 1 und 2.115 Tests im Fall von Aufgabe 2. Allerdings zeigt sich für Aufgabe 2, dass eine Gruppengröße von 10 bis 12 Personen pro Treatment für eine verlässliche Berechnung einer Kovarianzanalyse nicht ausreicht, da in diesen Fällen häufig die zugrunde liegende Modellmatrix singulär wird. Des Weiteren treten zunehmend numerische Instabilitäten auf. Diese äußern sich dergestalt, dass die geschätzten Mittelwerte weit außerhalb des zulässigen Wertebereichs liegen. So werden bspw. für Verhältnismaße, deren Wertebereich zwischen Null und Eins liegt, Mittelwerte von über zehn oder sogar negative Werte geschätzt. Somit sind in diesem Fall keine reliablen Aussagen möglich, weshalb an dieser Stelle nur die Ergebnisse für Aufgabe 1 berichtet werden. Aufgrund der immer noch verbleibenden Fülle der Daten, konzentriert sich die folgende

Tab. 6.26.: Übersicht über die Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses personenbezogener Störfaktoren auf die Benutzerleistung für $A1$ in SP_A. Die Tabelle stellt für jede Kovariate alle Leistungsmaße mit entfallenden (–) bzw. hinzukommenden (+) Effekten der Systemleistung (S), der Erwartung (E) oder ihre Interaktion (I) dar.

ID	Beschreibung	K04	K05	K06	K08	K10	K12
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	$-$S	$-$S	$-$S	$-$S	$-$S	$-$S
V17	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	-	-	-	$+I_{SE}$	-	-
V18	$\frac{\text{Anz. richtig rel. bew. Dok. (erste 10 Dok.)}}{\text{Anz. aufg. Dok. (erste 10 Dok.)}}$	-	-	+S	-	-	-
V19	$\frac{\text{Anz. richtig rel. bew. Dok. (erste 10 Dok.)}}{\text{Anz. rel. bew. Dok. (erste 10 Dok.)}}$	-	$-$S	-	-	-	-
V21	$\frac{\text{Anz. richtig rel. bew. Dok. (erste Suche)}}{\text{Anz. rel. bew. Dok. (erste Suche)}}$	$+I_{SE}$	$+I_{SE}$	$+I_{SE}$ +E	-	$+I_{SE}$	-
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	-	-	-	$-$S	-	$-$S
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	-	-	-	-	-	+S

Darstellung weiterhin ausschließlich auf solche Variablen, bei denen durch die Kovarianzanalyse signifikante Effekte hinzukommen oder verschwinden.

Die entsprechenden Ergebnisse sind getrennt nach Leistungsmaßen und Zufriedenheitsindikatoren in den Tabellen 6.26 und 6.27 aufgeführt. In den Zeilen der Tabelle stehen die abhängigen Variablen, in den Spalten hingegen die getesteten Kovariaten. Die Zellen geben Auskunft über das Hinzukommen (+) oder Verschwinden (−) eines Haupteffekts (S = System, E = Erwartung) oder einer Interaktion (I). So bedeutet das Kürzel $+I_{SE}$ bspw., dass ein signifikanter Interaktionseffekt zwischen System und Erwartung hinzukommt. Die zugehörigen Mittelwerte und Signifikanzniveaus sowie weiterführende Informationen bezüglich der zugrunde liegenden Stichprobengröße, der verwendeten Varianzanalyse (klassisch vs. robust) und der Stabilität des jeweiligen Effekts (eindeutig vs. tendenziell) können in Anhang C.5 in den Tabellen C.41 bis C.46 eingesehen werden.

Es zeigt sich, dass im Wesentlichen die in Abschnitt 6.4.3 und 6.4.5 beschriebenen Effekte auch unter Einbeziehung der betrachteten Kovariaten erhalten bleiben. Für die Benutzerleistung weisen sieben Variablen einen Kovarianzeffekt auf, wobei sechs der zwölf Kovariaten keinen Einfluss auf die Ergebnisse haben (K01, K02, K03, K07, K09 u. K11). In keinem Fall führt die Hinzunahme von Kovariaten zu einer qualitativen Änderung eines im Rahmen der Hauptanalyse signifikanten Haupteffekts. Für die Zufriedenheit hingegen weisen 13 Variablen einen Kovarianzeffekt auf. Hier hat nur die Muttersprache K03 keinen Einfluss auf die Ergebnisse.

Im Folgenden wird zunächst genauer auf die Ergebnisse im Kontext der Benutzerleistung eingegangen. Hier fallen drei der in der Hauptanalyse beschriebenen Effekte weg (B06, V19 u. V28/PCP). Kritisch zu betrachten sind besonders die Änderungen in Bezug auf B06, da diese in direktem Zusammenhang mit dem systemabhängigen Anpassungseffekt stehen. Allerdings handelt es sich hierbei nur um ein einzelnes Leistungsmaß, das sich darüber hinaus sehr speziell nur auf die letzte von den Probanden durchgeführte Suche bezieht. Ähnliches gilt für die Precisionmaße

Tab. 6.27.: Übersicht über Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses personenbezogener Störfaktoren auf die Benutzerzufriedenheit für $A1$ in SP_A. Die Tabelle stellt für jede Kovariate alle Zufriedenheitsindikatoren mit entfallenden (–) bzw. hinzukommenden (+) Effekten der Systemleistung (S), der Erwartung (E) oder ihre Interaktion (I) dar.

ID	Beschreibung	K01	K02	K04	K05	K06	K07	K08	K09	K10	K11	K12
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	-	-	+E	+E	+E	-	-	+E	+E	-	-
F09	Ist die Suchmaschine einfach zu bedienen?	−E	-	−E	−E	−E	-	−E	−E	-	-	−E
F11	Liefert die Suchmaschine aktuelle Information?	-	-	$+I_{SE}$ +S	$+I_{SE}$ +S	$+I_{SE}$ +S	-	$+I_{SE}$ +S	-	-	-	$+I_{SE}$ +S
F12	Ist die Suchmaschine erfolgreich?	-	-	$+I_{SE}$ +S	$+I_{SE}$ +S	$+I_{SE}$ +S	-	$+I_{SE}$ +S	-	$+I_{SE}$ +S	-	-
F14	Es war einfach, die Aufgabe zu bearbeiten.	-	-	-	-	-	-	-	-	-	+E	-
SK02-M	Inhalt	−E	-	-	−E	-	-	-	-	-	-	-
SK04-M	Suche	-	−E	-	-	-	-	-	-	-	-	-
SK05-F	Aufgabe	-	+E	-	-	+E	-	+E	-	-	-	+E
SK05-M	Aufgabe	-	-	-	-	+E	-	+E	-	-	-	+E
SK06-M	Eigenleistung	−E	−E	−E	−E	−E	−E	−E	−E	−E	-	-
SK10-F	Aufgabe	-	-	-	-	+E	-	-	-	-	-	-
SK11-M	Eigenleistung	-	-	-	-	-	+E	-	-	-	-	-
SK-C	Content (EUCS)	−E	−E	-	−E	-	-	-	-	−E	-	-

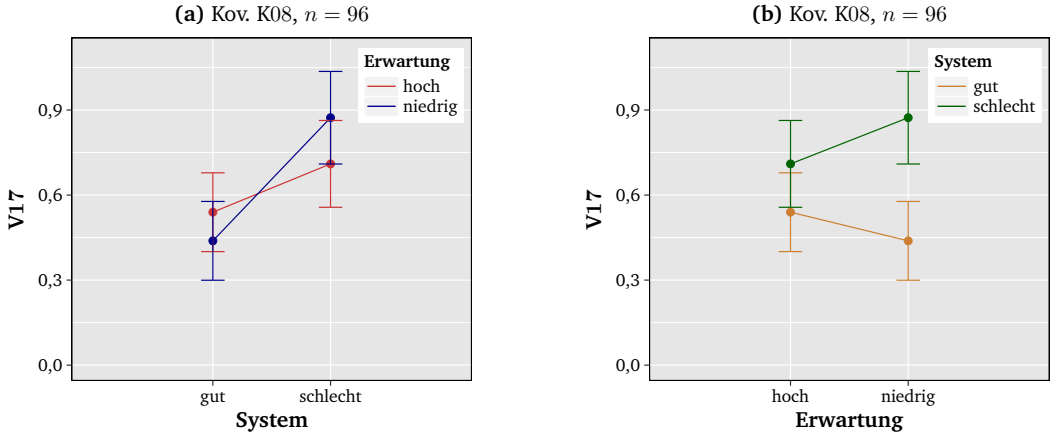


Abb. 6.17.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für den Anteil der richtig als irrelevant bewerteten Dokumente an allen als irrelevant bewerteten Dokumenten (V17) unter Berücksichtigung der Einschätzung des eigenen Suchmaschinenwissens (K08) als Kovariate. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bei Testpersonen mit hoher Erwartungshaltung ist der Einfluss der Systemleistung unterdrückt. Im Fall einer niedrigen Erwartungshaltung hingegen führt die bessere Systemleistung zu einem geringeren Anteil richtig als irrelevant bewerteter Dokumente. Dies deutet erneut auf strenger Relevanzkriterien bei der besseren Systemleistung hin, da etwa die Hälfte der als irrelevant bewerteten Dokumente von den Juroren als relevant eingestuft wurde. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

V19 und V28/PCP. Die Validität des Anpassungseffekts ist somit nicht gefährdet. Er wird weiterhin gestärkt durch den nun signifikanten Befund für V29, bei dem der Anteil der richtig als relevant erkannten Dokumente für das schlechtere System besser ausfällt. Interessant ist in diesem Zusammenhang, dass bei dem verwandten Benutzerleistungsmaß V18, bei dem nur die ersten zehn angezeigten Dokumente berücksichtigt werden, der systembedingte Anpassungseffekt noch nicht

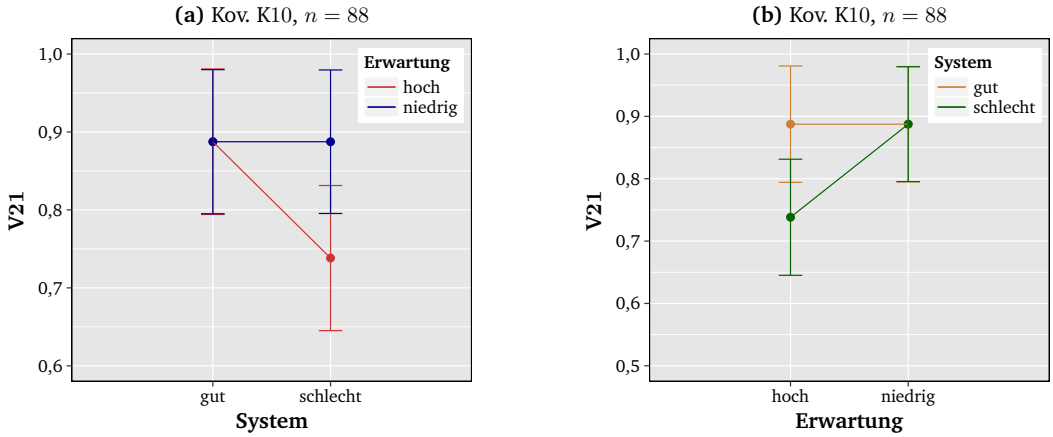


Abb. 6.18.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für den Anteil der richtig als relevant bewerteten Dokumente an allen als relevant bewerteten Dokumenten der ersten Suche (V21) unter Berücksichtigung der Suchmaschinennutzungsstunden (K10) als Kovariate. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bei der besseren Systemleistung findet sich kein Einfluss der Erwartungshaltung, während bei niedriger Systemleistung und hoher Erwartung der Anteil richtig als relevant identifizierter Dokumente sinkt, also auch irrelevante Dokumente als relevant akzeptiert werden. Umgekehrt ist eine systemleistungsbedingte Anpassung der Relevanzbewertung nur bei der hohen Erwartungshaltung feststellbar. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

sichtbar ist. Hier fällt stattdessen der Anteil richtig relevant bewerteter Dokumente bei dem besseren System höher aus. Dies könnte darauf hindeuten, dass die Probanden die Systemqualität erst wahrnehmen müssen, bevor der Effekt zum Tragen kommt. Allerdings ist zu beachten, dass in diesem Fall in Bezug auf alle aufgerufenen Dokumente anstelle der aufgerufenen relevanten Dokumente normalisiert wird.

Weiter präzisiert wird der Anpassungseffekt durch die hinzukommende Interaktion zwischen Systemleistung und Erwartungshaltung für V17, wenn die Selbsteinschätzung des Suchmaschinenwissens (K08) mit in die Betrachtung einbezogen wird. Wie Abbildung 6.17 zu entnehmen, ist es zwar nach wie vor der Fall, dass das bessere System zu einer strengeren Relevanzbewertung führt: Es werden mehr relevante Dokumente als irrelevant bewertet, was in der Konsequenz zu einem geringeren Anteil richtig als irrelevant bewerteter Dokumente führt. Jedoch hängt die Effektstärke nun von der Erwartungshaltung der Testpersonen ab: Bei niedriger Erwartungshaltung fällt der Unterschied zwischen den beiden Systemleistungen stärker aus. In der Kombination schlechtes System mit niedriger Erwartungshaltung bewerten die Testpersonen also noch weniger relevante Dokumente als irrelevant als in der Kombination schlechtes System mit hoher Erwartungshaltung.

Ähnliches gilt für die hinzukommende Interaktion in Bezug auf V21. Auch hier ermöglicht die Einbeziehung von Kovariaten eine detailliertere Einordnung des Effekts. Wie Abbildung 6.18 zu entnehmen ist, wird der bereits im Rahmen der Hauptauswertung beschriebene systembedingte Anpassungseffekt durch die Erwartungshaltung moderiert. Ein Unterschied zwischen hohem und niedrigem System tritt für die erste Suche im Gegensatz zu V17 jedoch ausschließlich in Verbindung mit einer hohen Erwartung der Testpersonen zu Tage, wohingegen eine niedrige Erwartung keinen Unterschied hervorruft. Einzig die Kombination aus niedrigem System und

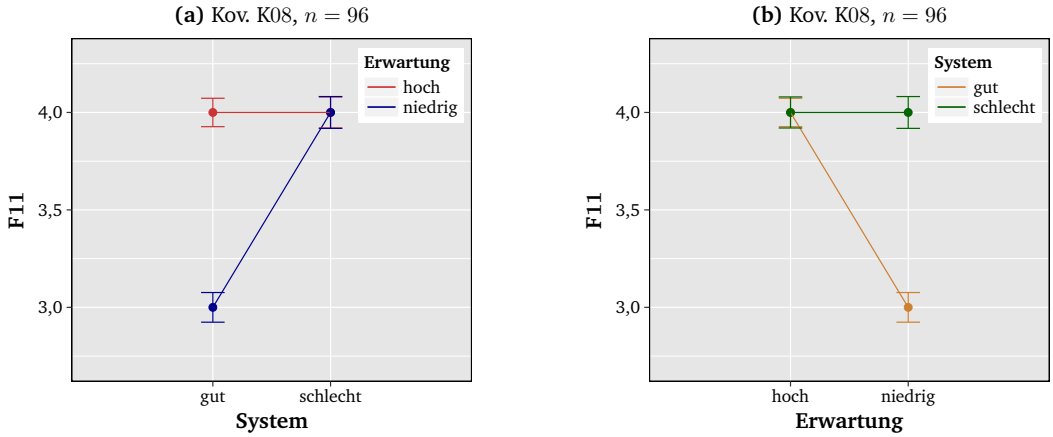


Abb. 6.19.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für das Zufriedenheitsitem: *Liefert die Suchmaschine aktuelle Information?* (F11) unter Berücksichtigung der Einschätzung des eigenen Suchmaschinenwissens (K08) als Kovariate. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bei geringer Systemleistung findet sich kein Einfluss der Erwartungshaltung, während bei besserer Systemleistung und niedriger Erwartung die Zufriedenheit geringer ausfällt. Umgekehrt ist eine systemleistungsbedingte Anpassung der Zufriedenheit nur bei der niedrigen Erwartungshaltung feststellbar. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

hoher Erwartung führt im Zuge der ersten Suche zu einer weniger strengen Relevanzbewertung. Umgekehrt lässt sich diese Wechselwirkung auch als systemabhängiger Erwartungseffekt interpretieren, bei dem im Kontext der geringen Systemleistung eine hohe Erwartungshaltung zu einem weniger restriktiven Bewertungsverhalten führt, bei dem mehr irrelevante Dokumente als relevant akzeptiert werden.

Im Folgenden werden die Ergebnisse der Kovarianzanalysen für die Benutzerzufriedenheit beschrieben. Neben einem Wegfall und einem Hinzukommen von Erwartungseffekten, was für fünf bzw. sechs Items der Fall ist, treten nun auch Interaktionseffekte zwischen System und Erwartung auf. Die Variablen, bei denen der Erwartungseinfluss verschwindet, können grob in zwei Gruppen unterteilt werden. Instabile Effekte, die bei der Hinzunahme vieler der betrachteten Kovariaten verschwinden (F09 u. SK06-M) und Variablen, bei denen der Erwartungseinfluss nur durch wenige Kovariaten annulliert wird (SK02-M, SK04-M u. SK-C). So ist es interessant, dass die beiden inhaltszentrierten Skalen SK02-M und SK-C beide durch das Herausrechnen des Alters (K01) und der Computernutzung (K05) den signifikanten Erwartungseinfluss verlieren. Für SK-C gilt dies auch in Bezug auf die Suchmaschinennutzung (K10). Eine mögliche Erklärung wäre, dass die Vertrautheit im Umgang mit Suchmaschinen zu Recherchezwecken die zuvor beobachteten Zufriedenheitsunterschiede hervorgerufen hat. Darüber hinaus finden sich geschlechtsspezifische Effekte für die Skalen SK04-M und SK-C. Da jedoch mit fünf Zufriedenheitsindikatoren nur eine geringe Anzahl von Effekten wegfällt, bleiben insgesamt die in der Hauptanalyse getroffenen Aussagen unberührt. Bei den zusätzlichen Erwartungseffekten bestätigt sich in allen Fällen der schon in der Hauptanalyse beobachtete Effekt des direkten Übergangs der induzierten Erwartung in die Benutzerzufriedenheit.

Die neu auftretenden Interaktionseffekte für die Zufriedenheitsitems F11 und F12 zeigen für

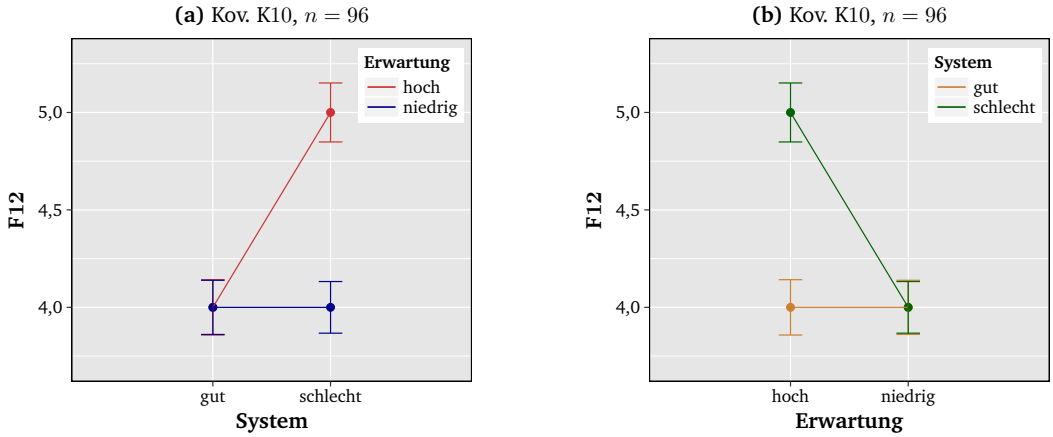


Abb. 6.20.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für das Zufriedenheitsitem: *Ist die Suchmaschine erfolgreich?* (F12) unter Berücksichtigung der Suchmaschinennutzungsstunden (K10) als Kovariate. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bei hoher Systemleistung findet sich kein Einfluss der Erwartungshaltung, während bei niedriger Systemleistung und hoher Erwartung die Zufriedenheit höher ausfällt. Umgekehrt ist eine systemleistungsbedingte Anpassung der Zufriedenheit nur bei der hoher Erwartungshaltung feststellbar. Fehlerbalken kennzeichnen $95\,\%$-Konfidenzintervalle der Gruppenmittelwerte.

alle relevanten Kovariaten ein qualitativ ähnliches Verhalten. Dabei können die Interaktionsdiagramme in beiden Fällen die Vorhersagen des C/D-Paradigmas nicht bestätigen (vgl. Abb. 6.19 u. 6.20). Im Falle von F11 sind die Probanden mit dem besseren System und der niedrigen Erwartungshaltung signifikant unzufriedener als Teilnehmer der übrigen Untersuchungsgruppen, während das C/D-Paradigma aufgrund einer positiven Diskonfirmation gerade das gegenteilige Verhalten vorhersagen würde. Für F12 hingegen geben Teilnehmer des schlechteren Systems mit der hohen Erwartungshaltung ein signifikant höheres Zufriedenheitsurteil ab, was in gleicher Weise den Aussagen des C/D-Paradigmas zuwider läuft. Das Bewertungsverhalten in Bezug auf F12 steht dabei jedoch im Einklang mit der zuvor beschriebenen Interaktion für das Leistungsmaß V21 (vgl. Abb. 6.18). Hier fällt der Anteil richtig als relevant bewerteter Dokument für die Kombination hohe Erwartungshaltung mit schlechtem System geringer als für die anderen drei Treatmentgruppen aus, was auf weniger strenge Relevanzkriterien in dieser Untersuchungsgruppe hindeutet. Diese Probanden nehmen das System also als positiver wahr, da sie mehr Dokumente als relevant akzeptieren und somit erscheint es plausibel, dass auch ihr Zufriedenheitsurteil positiver ausfällt.

Für F11 hingegen zeigt sich ein Einfluss der Erwartungshaltung ausschließlich im Kontext des besseren Suchsystems bzw. umgekehrt ein Einfluss der Systemgüte nur im Rahmen der niedrigen Erwartungshaltung (vgl. Abb. 6.19). Hier ist zunächst anzumerken, dass in Bezug auf die Aktualität der präsentierten Dokumente kein Unterschied zwischen den beiden Suchsystemen besteht, da beide auf dieselbe Datenbasis zugreifen. Allerdings scheint sich in dem hier beobachteten Zufriedenheitsurteil tendenziell das Verhalten in Bezug auf Leistungsmaß V17 widerzuspiegeln, da die Untersuchungsgruppe mit dem strengsten Bewertungsverhalten (bessere Systemqualität/niedrige Erwartungshaltung) und damit dem negativsten Eindruck von der Sys-

temqualität das pessimistischste Zufriedenheitsurteil fällt. Dabei ist jedoch zu berücksichtigen, dass die Unterschiede in Bezug auf die Zufriedenheit weit größer ausfallen, als dies im Rahmen der Benutzerleistung der Fall ist.

Zusammenfassend lässt sich sagen, dass die wesentlichen Befunde der Hauptanalyse auch unter Einbeziehung personenbezogener Störfaktoren Bestand haben. Dies gilt sowohl mit Blick auf die Benutzerleistungsmaße und den beschriebenen Anpassungseffekt der Relevanzbewertung, als auch für die Benutzerzufriedenheit. Es wäre wünschenswert gewesen, diese Analyse auch für Aufgabe 2 vorzunehmen, jedoch verhindern hier die geringen Fallzahlen eine reliable Auswertung. Zwar entfallen unter der Einbeziehung von Kovariaten vereinzelt signifikante Haupteffekte, die im Rahmen der Hauptuntersuchung beschriebenen Ergebnisse beruhen allerdings stets auf mehreren Indikatoren. Dies zeigt einmal mehr, dass eine breit angelegte Erfassung der Benutzerleistung einen Vorteil des gewählten Untersuchungsdesigns darstellt. Darüber hinaus erlauben die Kovarianzanalysen einen etwas detaillierteren Blick auf einige Befunde indem nun auch Wechselwirkungen zwischen System und Erwartung signifikant werden.

6.5. Fazit: Experiment 2

Den Ausgangspunkt des zweiten Experiments bildet der Anspruch, aufbauend auf den Ergebnissen des ersten Experiments, den Einfluss von Systemleistung und Benutzererwartungen während des Suchprozesses auf das Leistungs- und Zufriedenheitsverhalten der Benutzer genauer zu untersuchen. Während die Testpersonen in Experiment 1 nur ein Treatment erhalten, wird die Erwartungsmanipulation diesmal über einen Vergleich von zwei unterschiedlich guten Systemen realisiert. Dazu wird die schriftliche Testinstruktion durch ein kurzes Einführungsvideo ersetzt und die Versuchspersonen müssen am Ende des Benutzertests eine Kontrollfrage beantworten, die überprüfen soll, ob die Erwartungsmanipulation erfolgreich ist.

Einige methodische Herausforderungen interaktiver Retrievalexperimente können im Rahmen des zweiten Experiments besser gelöst werden. So lässt sich auf der einen Seite eine erhebliche Verbesserung in Bezug auf den Realitätsgrad des experimentellen Designs feststellen. Insbesondere wird mit der freien Suchbegriffswahl eine entscheidende Verbesserung in Bezug auf ein natürlicheres Sucherlebnis erzielt. Auf der anderen Seite sind mit diesen Verbesserungen auch eine erhöhte Komplexität sowie eine geringere Vorhersagbarkeit und Kontrolle verbunden, da sich aus dem veränderten Untersuchungsablauf weitere Störvariaben ergeben können. So stellt sich in diesem Zusammenhang bspw. die Frage, wie die Benutzerleistung gemessen werden soll, wenn jede Testperson unterschiedlich viele Suchanfragen stellt und dafür unterschiedliche Suchergebnislisten erhält. Während die Suchleistung nämlich im ersten Experiment anhand von nur zwei Trefferlisten pro Suchthema verglichen werden muss, bilden im zweiten Experiment zufällig generierte Trefferlisten die Grundlage für den Benutzerleistungsvergleich (vgl. Abschn. 6.3.2). Ein weiterer Aspekt, der zur Komplexität von Experiment 2 beiträgt, ist der Systemvergleich. Dieser macht es erforderlich, die Nutzungserfahrung des ersten Systems bei der Analyse der zweiten Aufgabe als weiteren Faktor einzubeziehen (vgl. Abschn. 6.4.3).

Die Ergebnisse aus Experiment 1 bezüglich der Kontextabhängigkeit des Bewertungsverhaltens können vollumfänglich bestätigt werden. Auch im zweiten Experiment ruft eine höhere Systemleistung im Kontext precisionorientierter Leistungsmaße strengere Bewertungsmaßstäbe bei der

Relevanzbewertung der Suchergebnisse hervor. Gleiches gilt für die Kompensation der Systemunterschiede bei recallorientierten Maßen. Auch hier sind die Probanden wie im ersten Experiment dazu in der Lage, ihr Informationsbedürfnis auch bei geringerer Systemqualität zu befriedigen. Dies kann als Bestätigung der Ergebnisse von Al-Maskari et al. (2008b) angesehen werden, die einen ähnlichen Kompensationseffekt für einen vergleichbaren relativen Systemunterschied von $30\,\%$ beobachten. Zusätzlich kann ein leichter Effekt der Erwartungshaltung auf die Benutzerleistung festgestellt werden, indem eine höhere Erwartungshaltung zu einer besseren PCP führt. Insgesamt sprechen die Ergebnisse der Benutzerleistung somit für einen differierenden Einfluss der Systemleistung auf unterschiedliche Aspekte des Suchverhaltens, während der Einfluss der Erwartungshaltung noch weiter untersucht werden muss.

Hinsichtlich der Zufriedenheit der Testpersonen kann im Rahmen von Experiment 2 darüber hinaus ein signifikanter Effekt der Erwartungshaltung nachgewiesen werden. Es zeigt sich, dass der Einfluss der Erwartungshaltung stark vom Erhebungszeitpunkt der Zufriedenheitsreaktion abhängt. Während die Ergebnisse der ersten Aufgabe noch einen Haupteffekt der Erwartung zeigen, scheint dieser mit zunehmender eigener Erfahrung nachzulassen, sodass im Rahmen der zweiten Aufgabe nur noch die Systemleistung signifikant ist. Damit weisen die Ergebnisse auf eine dynamische Abhängigkeit der Zufriedenheit hin, dessen Analyse zentraler Bestandteil des dritten Experiments sein wird. Betrachtet man schließlich den tatsächlichen Effekt, den die Erwartungshaltung auf die Benutzerzufriedenheit ausübt, so zeigt sich, dass die Vorhersage des C/D-Paradigmas, dass Benutzer mit niedrigeren Erwartungen insgesamt betrachtet zufriedener sind, hier erneut nicht bestätigt werden kann. Vielmehr scheint es so zu sein, dass hohe Erwartungen im IR-Kontext sowohl zu einer besseren Leistung (wie anhand des signifikanten Effekts bei der PCP zu sehen) als auch zu einer erhöhten Zufriedenheit beitragen. Aus forschungstheoretischer Perspektive könnte dies bedeuten, dass eine Übertragung des C/D-Paradigmas auf den Kontext der Informationssuche im Internet nicht möglich ist, da das Zufriedenheitsurteil in diesem Fall zusätzlich durch die Wahrnehmung der eigenen Leistung beeinflusst wird. Auch diesen Zusammenhang gilt es, in Folgestudien weiter zu untersuchen, da der beobachtete Effekt der Erwartungshaltung auf die Benutzerleistung im zweiten Experiment nur gering ausfällt. Weiterhin zeigt sich, dass alle beobachteten Effekte stabil in Bezug auf eine Vielzahl von personen- und untersuchungsabhängigen Störfaktoren sind.

Um die angesprochenen Schwächen des vorliegenden Untersuchungsdesigns zu überwinden, sollten für Folgestudien vor allem folgende Änderungen in Betracht gezogen werden: Um die Komplexität des Studiendesigns sinnvoll zu reduzieren, könnte der Systemvergleich erneut auf die Nutzung eines Systems beschränkt und den Probanden lediglich mitgeteilt werden, dass das Ziel der Untersuchung darin besteht, zwei Systeme zu vergleichen. Um die dynamische Entwicklung der Qualitätswahrnehmung genauer untersuchen zu können, ist es darüber hinaus erforderlich, dass die Probanden mehrere Suchaufgaben mit ähnlichen Wahrnehmungs- und Handlungsanforderungen ausführen und die Zufriedenheit der Testteilnehmer nach jeder einzelnen Aufgabe erhoben wird.

7. Experiment 3: Dynamische Entwicklung der wahrgenommenen Retrievalqualität

Das dritte im Rahmen dieser Arbeit durchgeführte Experiment erweitert den Fokus der Untersuchung um die Analyse prozessualer Aspekte des Erwartungs-Wahrnehmungs-Vergleichs. Folgende in der dritten Forschungsfrage (vgl. Abschn. 1.2) konkretisierte Problemstellungen stehen dabei im Mittelpunkt: die Exploration der Suchergebnisliste, die Relevanzwahrnehmung im Suchverlauf sowie die nachträgliche Anpassung der Benutzererwartung. Bei der Entwicklung des Untersuchungsdesigns wird Wert darauf gelegt, dass sowohl Selbstauskunfts- als auch Verhaltensmaße im zeitlichen Verlauf untersucht werden können, d.h. die Veränderung der Wahrnehmung in ihrem prozessualen Kontext erfasst werden kann. Während also in den ersten beiden Experimenten die Erwartungshaltung als unabhängige und zeitlich konstante Variable in die Betrachtung einfließt, wird nun auch die aus der Erfahrung der vorherigen Aufgabe resultierende Erwartungsanpassung berücksichtigt. Die Darstellung folgt dabei zur besseren Vergleichbarkeit der Struktur der beiden vorangehenden Kapitel.

7.1. Untersuchungsziel

Wenngleich in den ersten beiden Experimenten ein Zusammenhang zwischen Benutzererwartungen und Benutzerzufriedenheit nachgewiesen werden kann, zeigt sich noch kein einheitliches Bild bezüglich der spezifischen Wirkungsmechanismen der Erwartungshaltung. Während das C/D-Paradigma im ersten Experiment in der Tendenz sichtbar, jedoch nicht signifikant ist, zeigt die Erwartungsmanipulation im zweiten Experiment zwar teilweise einen signifikanten Einfluss auf das Zufriedenheitsurteil, die Wirkungsrichtung der Erwartungshaltung entspricht jedoch nicht den Vorhersagen des C/D-Paradigmas, wonach Benutzer mit niedrigeren Erwartungen insgesamt betrachtet zufriedener sein sollten als Benutzer mit einer höheren Erwartungshaltung. Verschiedene in Abschnitt 6.5 diskutierte Anhaltspunkte lassen vermuten, dass diese uneinheitlichen Ergebnisse in den unterschiedlichen Untersuchungsanordnungen der ersten beiden Experimente begründet sein könnten.

Im dritten Experiment werden deshalb die Vorteile der ersten beiden Experimente in Bezug auf die Manipulation der Erwartungshaltung zusammengeführt. Um eine ähnliche Manipulationsstärke wie im zweiten Experiment zu erreichen, wird den Testpersonen weiterhin der Vergleich von zwei verschiedenen Suchmaschinen suggeriert. Um allerdings die Komplexität des Studiendesigns zu reduzieren, wird der Benutzertest pro Versuchsperson nur mit einem der Systeme durchgeführt. Um eine bessere Generalisierbarkeit zu gewährleisten und Topic- und Lerneffekte zu kontrollieren, sollen hingegen, wie im ersten Experiment geschehen, drei unterschiedliche Suchaufgaben bearbeitet werden. Zur Ermittlung möglichst realistischer Zufriedenheitsurteile wird die Zufriedenheit der Testteilnehmer außerdem nach jeder einzelnen Aufgabe erhoben. Die-

ses Vorgehen ermöglicht zugleich die Untersuchung der dynamischen Abhängigkeit der Benutzerzufriedenheit. Ein weiteres Anliegen bei der Erarbeitung des Untersuchungsdesigns besteht darin, diesmal auch die Relevanzurteile der Testpersonen genauer zu analysieren. Diesbezüglich wird in Anlehnung an den Stand der Forschung (4.2.2.1) davon ausgegangen, dass eine binäre Skala nicht ausreicht, um die Erscheinungsvielfalt menschlicher Relevanzurteile vollständig abzubilden, sodass den Testpersonen diesmal eine 8-stufige Bewertungsskala zur Verfügung steht.

7.2. Forschungsleitende Hypothesen

Die dritte Nutzerstudie dient zum einen der Überprüfung der im Rahmen der ersten beiden Experimente erhaltenen Ergebnisse, erweitert den dort verfolgten Ansatz aber insbesondere um eine Betrachtung der dynamischen Entwicklung des Nutzerverhaltens. Dazu wird zunächst die Frage nach der Gültigkeit des C/D-Paradigma erneut aufgegriffen. Zwar lassen die Ergebnisse der zweiten Studie zusammen mit dem nicht-signifikanten Erwartungseinfluss des ersten Experiments bereits vermuten, dass eine direkte Übertragung des C/D-Paradigmas auf den IR-Bereich nicht möglich ist. Um jedoch sicherzustellen, dass der beobachtete Erwartungseffekt keine Folge des komplexen Untersuchungsdesigns des zweiten Experiments darstellt, wird die erste Forschungshypothese weiterhin aufrechterhalten.

H1: Die Zufriedenheit der Benutzer wird durch ihre Erwartungshaltung und die Systemgüte gemäß den Annahmen des C/D-Paradigmas beeinflusst.

Die nächsten beiden Hypothesen (H2 u. H3) betreffen erneut den Zusammenhang von System- und Benutzerleistung. Beide Hypothesen können durch die Ergebnisse des zweiten Experiments bestätigt werden. Dies betrifft zum einen den nicht signifikanten Einfluss der Systemgüte auf recallorientierte Benutzerleistungsmaße und zum anderen den systembedingten Anpassungseffekt der Relevanzwahrnehmung in Bezug auf precisionorientierte Leistungsindikatoren. Durch die nochmalige Überprüfung dieser Ergebnisse auf Basis einer weiteren Stichprobe und unter Verwendung neuer Suchthemen soll nun die Validität der Ergebnisse erhöht werden.

H2: Bei recallorientierten Leistungsmaßen können Benutzer Unterschiede in der Systemgüte kompensieren.

H3: Bei precisionorientierten Leistungsmaßen passen Benutzer ihre Relevanzdefinition der Systemgüte an.

In Bezug auf die Erwartungshaltung lassen sich im Hinblick auf die beiden vorangegangenen Nutzerstudien zwei Befunde festhalten: Zum einen führt eine Zufriedenheitsmessung im Anschluss an drei bearbeitete Aufgaben zu keinem messbaren Einfluss der Erwartungshaltung (Experiment 1), zum anderen zeigen sich im zweiten Experiment signifikante Erwartungseinflüsse primär in Bezug auf die erste Aufgabe. Eine plausible Erklärung dieser Beobachtungen ergibt sich aus der in Abschnitt 2.1.1.3 beschriebenen Studie von Szajna und Scamell (1993), wonach mit einem Nachlassen unrealistischer Erwartungen im Rahmen der wiederholten Nutzung eines Informationssystems auszugehen ist. Diese Beobachtung ist Anlass im dritten Experiment die dynamische Wahrnehmung der Retrievalqualität genauer zu untersuchen. Es soll also gezeigt

werden, dass unrealistische Benutzererwartungen mit der Zeit an Bedeutung verlieren. Entsprechend lautet die vierte Forschungshypothese des dritten Experiments:

H4: Im Verlauf der Systemnutzung passen Benutzer ihre unrealistischen Erwartungen der Systemgüte an.

7.3. Methode

Um im dritten Experiment die Qualitätswahrnehmung im zeitlichen Verlauf analysieren zu können, wird ein zweifaktorielles Design mit Messwiederholung auf beiden Faktoren entworfen. Wie in den vorangegangenen beiden Experimenten stellen die Erwartungshaltung und die Systemqualität die unabhängigen Variablen der Untersuchung dar. Beide Variablen besitzen eine niedrige und eine hohe Ausprägung, sodass sich der in Abbildung 7.1 dargestellte Versuchsplan mit vier Versuchsgruppen ergibt. Die Zuordnung der Testteilnehmer erfolgt dabei randomisiert in einem Doppelblindverfahren. Dies bedeutet, dass weder der Testleiter noch die Testpersonen wissen, welcher Versuchsgruppe sie zugeordnet werden (vgl. Abschn. 4.2.3.2).

		System					
		gut			schlecht		
Erwartung	niedrig	Gruppe 1			Gruppe 2		
		A_1	A_2	A_3	A_1	A_2	A_3
	hoch	Gruppe 3			Gruppe 4		
		A_1	A_2	A_3	A_1	A_2	A_3

Abb. 7.1.: Versuchsplan des dritten Experiments. In dem gewählten Between-Subjects-Design führt die zweifache Abstufung der beiden Faktoren Systemgüte und Erwartungshaltung zu den dargestellten vier Untersuchungsgruppen. Jede Testperson bearbeitet drei Suchaufgaben, für die jeweils sowohl die Benutzerzufriedenheit als auch die Benutzerleistung gemessen wird.

Damit entspricht das Untersuchungsdesign des dritten Experiments einer Zusammenführung und Präzisierung der ersten beiden Experimente. Zwar wird allen Teilnehmern wie im zweiten Experiment weiterhin der Vergleich von zwei Systemen suggeriert, jedoch bearbeiten die Probanden alle drei Suchaufgaben (A1 bis A3) wie im ersten Experiment bei konstant bleibender Systemleistung. Die Messung der wahrgenommenen Retrievalqualität erfolgt dabei im Anschluss an jede einzelne Aufgabe. Die angestrebte Stichprobengröße wird auf 25 Probanden pro Versuchsgruppe, also insgesamt 100 Teilnehmer festgelegt.

7.3.1. Manipulation der unabhängigen Variablen

Die Erwartungshaltung der Testpersonen und die Retrievalqualität des Suchsystems bilden auch im dritten Experiment die unabhängigen Variablen. In Bezug auf die Systemqualität werden dabei im Sinne der Vergleichbarkeit der Ergebnisse die bereits für die ersten beiden Experimente gewählten Systemleistungsunterschiede übernommen (vgl. Abschn. 6.3.1). Wie im Rahmen der

forschungsleitenden Hypothesen in Abschnitt 7.2 bereits angedeutet, kommt im dritten Experiment für die Jurorenurteile jedoch eine feiner abgestufte Relevanzskala mit den Kategorien relevant, eher relevant, eher irrelevant und irrelevant zum Einsatz (vgl. Abschn. 7.3.4). Die tatsächliche Manipulation der Systemleistung hingegen erfolgt analog zu den ersten beiden Nutzerstudien anhand künstlich erzeugter Ergebnislisten. Die im zweiten Experiment eingeführte Möglichkeit der iterativen Suche wird beibehalten, sodass die in Abschnitt 6.3.1 beschriebenen Rankinglisten im dritten Experiment wiederverwendet werden können, was zusätzlich die Vergleichbarkeit der Ergebnisse erhöht. Die tatsächliche Zuteilung der Dokumente erfolgt erneut in Echtzeit parallel zu den Aktionen der Testpersonen. Grundlage der Rankings bilden wiederum die Relevanzurteile der Juroren, die diesmal anhand der beschriebenen 4-stufigen, kategorialen Relevanzskala erfasst werden (vgl. Abschn. 7.3.4). Bei der randomisierten Verteilung der Dokumente auf die Rankingplätze werden den relevanten Positionen jeweils zufällig zur Hälfte relevante und eher relevanten Dokumenten zugeordnet. Das analoge Vorgehen erfolgt zur Besetzung der irrelevanten Positionen mit irrelevanten und eher irrelevanten Dokumenten.

Die Manipulation der Erwartungshaltung erfolgt erneut während der Testeinführung. Wie im zweiten Experiment wird den Testpersonen dazu ein Einführungsvideo gezeigt, das sicherstellen soll, dass alle Testpersonen identische Vorinformationen erhalten. Um die Manipulation der Erwartungshaltung weiter zu verstärken, wird dabei auf eine audiovisuelle Darbietung der Informationen zurückgegriffen, bei dem eine den Instruktionstext enthaltende Powerpointpräsentation mit einem entsprechenden Audiofile kombiniert wird (vgl. Abschn. 4.2.1.2). Anstelle eines fiktiven Testszenarios wird die Manipulation der Erwartungshaltung diesmal in einen konkreten Forschungszusammenhang integriert. Zu diesem Zweck werden alle Teilnehmer zu Beginn des Experiments darüber informiert, dass aktuelle Studien zum Informationssuchverhalten ergeben haben, dass Benutzer Qualitätsunterschiede zwischen unterschiedlichen Suchmaschinen kompensieren können. Den Testteilnehmern wird mitgeteilt, dass das Ziel des Benutzertests darin besteht, diesen Sachverhalt näher zu untersuchen, indem die eine Hälfte der Teilnehmer während des Test mit einem besseren und die andere mit einem schlechteren System arbeiten werde. Das Ziel besteht darin, den Probanden so eine realistischere Vorstellung von den Hintergründen des Experiments zu geben und so dazu beizutragen, dass das Versuchsdesign insgesamt von den Probanden als noch realistischer empfunden wird. Um jedoch auszuschließen, dass die interne Validität der Untersuchung durch das Untersuchungsdesign verfälscht wird, indem die Testpersonen bspw. versuchen sich hypothesenkonform zu verhalten, wird der Begriff der *Kompensation* bei der Formulierung der Erwartungsmanipulation gezielt vermieden. Stattdessen wird eine Formulierung gewählt, in der erwähnt wird, dass aktuelle Studien zur Suchmaschinennutzung gezeigt haben, dass die Suchmaschinenleistung keinen großen Einfluss auf den Sucherfolg der Benutzer hat und Benutzer auch mit unterschiedlich guten Suchmaschinen vergleichbare Ergebnisse erreichen. Die vollständige Testinstruktion ist in Anhang D.1 nachzulesen. Zusammenfassend lässt sich an dieser Stelle also festhalten, dass den Testpersonen im dritten Experiment erneut der Vergleich von zwei Systemen suggeriert wird, da dieses Designelement dazu beizutragen scheint, dass den Probanden die Einnahme der jeweiligen Erwartungshaltung erleichtert wird. Um jedoch gleichzeitig die Komplexität des Studiendesigns zu reduzieren, wird der Benutzertest diesmal pro Versuchsperson mit nur einem der beiden Systeme durchgeführt.

Des Weiteren wird im dritten Experiment erneut mit Hilfe einer Kontrollfrage am Ende der Tests überprüft, ob die Manipulation erfolgreich ist. Sie lautet: Wissen Sie noch, welche der beiden Suchmaschinen Sie verwendet haben? Als mögliche Antworten können die Probanden zwischen der besseren und der schlechteren Suchmaschine auswählen oder die Frage mit *weiß nicht* beantworten. Durch die Vorgabe einer expliziten weiß-nicht-Antwortmöglichkeit soll vermieden werden, dass Teilnehmer eine der beiden Antwortoptionen raten. Neben dieser abschließenden Kontrollfrage, wird auch eine mögliche, aus der Interaktion mit dem Testsystem resultierende, dynamische Erwartungsänderung erfasst, indem die im Zuge des zweiten Experiments bereits beschriebenen Frageitems von Szajna und Scamell (1993) im Anschluss an jede Suchaufgabe abgefragt werden (vgl. Abschn. 6.3.1). Darüber hinaus wird im dritten Experiment der von den Probanden erwartete eigene Sucherfolg vor Bearbeitung der Aufgaben erhoben.

7.3.2. Operationalisierung der abhängigen Variablen

Ein wesentlicher Unterschied zwischen zweitem und drittem Experiment besteht darin, dass die Testpersonen im dritten Experiment nicht mehr nur zwei, sondern drei Suchaufgaben bearbeiten und folglich den Zufriedenheitsfragebogen dreimal beantworten müssen. Dies führt zu einem Zielkonflikt zwischen einer möglichst breiten und lückenlosen Datenbasis auf der einen und methodischen Idealvorstellungen, nach welchen die Probanden weder kognitiv noch zeitlich zu überlasten sind, auf der anderen Seite. Aus diesem Grund werden im dritten Experiment nicht alle 26 Items aus dem zweiten Experiment in den Fragebogen aufgenommen. Verzichtet wird auf die Items F21 (*Ich hatte genügend Zeit, um eine effektive Suche durchzuführen.*) und F15 (*Es war einfach, zu dem Thema zu suchen.*). Begründet werden kann der Verzicht auf F21 damit, dass dieses Item sich in der Itemanalyse als nur mäßig reliabel und valide herausgestellt hat (vgl. Abschn. 6.4.4.1). F15 wird ausgeschlossen, da die Vertrautheit mit dem Suchthema durch die Items F14 (*Es war einfach, die Aufgabe zu bearbeiten*) und F20 (*Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.*) bereits gut abgedeckt wird. Im Zuge dieser Reduzierung finden auch die entsprechenden von diesen beiden Frageitems abhängigen Skalen SK05 und SK06 keine Berücksichtigung bei der Auswertung der Benutzerzufriedenheit.

Eine weitere Änderung betrifft die zur Verfügung stehenden Merkmalsausprägungen der eingesetzten Zufriedenheitsitems. Während die 13 Items des EUCS-Instruments und die 11 selbst entwickelten Zusatzitems im zweiten Experiment anhand unterschiedlich Skalenbreiten (5- vs. 7-stufige Skala) einzuschätzen sind, werden die Antwortkategorien im dritten Experiment zu einer 5-stufigen Skala vereinheitlicht. Dieser Maßnahme liegt vor allem die pragmatische Überlegung zugrunde, dass der unmittelbare Vergleich der beiden Fragebogenteile im dritten Experiment erleichtert werden soll. Schulz et al. (2006, S. 88) weisen in diesem Zusammenhang auf die Gefahr hin, dass die Verwendung unterschiedlicher Skalenbreiten möglicherweise statistische Implikationen hat, die nicht inhaltlich interpretiert werden dürfen. So kann sich bspw. das Korrelationsverhalten der einzelnen Zufriedenheitsitems in Abhängigkeit von der Skalenbreite unterscheiden. In Bezug auf die Möglichkeit, unterschiedliche Skalen auf eine auf der prozentualen Zustimmung der Teilnehmer beruhende Einheitsskala zu transformieren, geben Schulz et al. (ebd.) zu bedenken, dass die Tendenz extreme Urteile zu vermeiden (vgl. Abschn. 4.2.3.2) je nach Originalskalierung zu unterschiedlichen Ergebnissen führt. So entspricht auf dieser Einheitsskala eine 4 der 5-stufigen Skala einem prozentualen Wert von 75, eine 6 der 7-stufigen Skala

hingegen einem Wert von $83{,}3$. Nach inhaltlicher Fertigstellung findet die praktische Umsetzung des Fragebogens in LimeSurvey (vgl. Abschn. 7.3.5) statt, sodass die Fragebögen diesmal nicht auf Papier sondern direkt am Computer beantwortet werden können. Dies ermöglicht eine zeitnahe Auswertung der Daten. Außerdem werden Eingabefehler bei der manuellen Datenerfassung vermieden.

Die Beurteilung der Suchleistung der Testpersonen geschieht wiederum mit Hilfe der in den ersten beiden Experimenten eingesetzten Leistungsmaße (vgl. Abschn. 5.3.2 u. 6.3.2). Wie bereits angedeutet, werden sowohl die Relevanzurteile der Juroren (vgl. Abschn. 7.3.4), als auch die Relevanzurteile der Probanden diesmal anhand einer mehrstufigen Skala erfasst. Dabei wird für die Jurorenbewertung auf eine von Sormunen (2000) eingeführte 4-stufige Relvanzklassifikation mit klar definierten Bewertungskategorien zurückgegriffen, um eine möglichst konsistente Korpusbewertung zu erreichen (vgl. Abschn. 4.1.3.3). Im Fall der Relevanzwahrnehmung durch die Probanden steht hingegen eine möglichst intuitive Verständlichkeit der Skala im Mittelpunkt, die gleichzeitig jedoch einen ausreichenden Differenzierungsgrad bzgl. der Antwortkategorien bietet. Unter Berücksichtigung der Ergebnisse von Tang et al. (1999), die die optimale Anzahl von Antwortkategorien zwischen 7 und 8 Antwortstufen verorten, wird deshalb zur Messung der subjektiven Relevanz eine 8-stufige Skala mit Werten zwischen 1 (= irrelevant) und 8 (= relevant) gewählt (vgl. Abschn. 4.2.2.1). Dabei ermöglicht die Verwendung einer 8-stufigen Skala zum einen einen direkten Vergleich der Relevanzurteile der Probanden mit den anhand einer 4-stufigen Skala erfassten Urteilen der Juroren. Zum anderen können beide Skalen dank einer geraden Anzahl an Antwortmöglichkeiten zu gleichen Teilen in eine binäre Skala umkodiert werden. In Bezug auf die Beurteilung der Suchleistung erlaubt dies, jeweils zwei Varianten der in Abschnitt 6.3.2 beschriebenen Leistungsmaße zu betrachten: Einerseits können im Sinne einer konservativen Auswertung der Suchleistung ausschließlich die als *relevant* bewerteten Dokumente (Skalenstufen $7-8$) bei der Berechnung der als richtig relevant bewerteten Dokumente mit einbezogen werden. Andererseits kann eine weniger restriktive Betrachtungsweise gewählt werden bei der sowohl die *relevanten* als auch die *eher relevanten* Dokumente (Skalenstufen $5-8$) beitragen. Insgesamt kommen damit 126 neue Leistungsindikatoren zu den bestehenden 86 Benutzerleistungsmaßen hinzu, die im Einzelnen in Anhang D.3 aufgeführt sind.

Abschließend lässt sich festhalten, dass das Ziel einer möglichst umfassenden Abdeckung von Zufriedenheits- und Leistungsaspekten auch im dritten Experiment erreicht wird. Insbesondere die verbesserte Relevanzmessung erlaubt dieses Mal eine noch detailliertere Erfassung des objektiven Sucherfolgs.

7.3.3. Umgang mit Störvariablen

In Bezug auf personenbezogene Störvariablen, deren Auftreten die Interpretation der Ergebnisse erschweren kann, wird im dritten Experiment erneut von einer möglichen Beeinflussung der abhängigen Variablen durch zusätzliche Störvariablen ausgegangen. Im Wesentlichen werden dabei die bereits im zweiten Experiment angewendeten Kontrollstrategien beibehalten. Die Rekrutierung der Untersuchungsteilnehmer konzentriert sich wiederum auf die Altersgruppe Studierender. Da das dritte Experiment als Doppelblindversuch (vgl. Abschn. 7.3) durchgeführt wird, ist eine geschlechtshomogene Randomisierung der Untersuchungsteilnehmer diesmal jedoch nicht ohne Weiteres praktikabel. Gleichermaßen wird die Muttersprache der Probanden,

wie schon in den ersten beiden Experimenten, nicht statistisch kontrolliert, sondern per Fragebogen erfasst, sodass ein entsprechender Einfluss ggf. mit Hilfe einer Kovarianzanalyse aus den Daten herausgerechnet werden kann.

Um die für die Teilnahme erforderliche Zeit in vertretbaren Grenzen zu halten, ist es weiterhin notwendig, eine Auswahl unter den prinzipiell möglichen Kontrollstrategien zu treffen. Aus diesem Grund wird das Erfahrungswissen der Testpersonen im dritten Experiment nicht, wie im zweiten Experiment geschehen, anhand zweier zusätzlicher Wissenstests erfasst. Stattdessen wird davon ausgegangen, dass die mit Hilfe der Randomisierung erzielte Ausschaltung von Störgrößen ausreichend ist. Mit Blick auf die Vergleichbarkeit mit den ersten beiden Experimenten bleiben die im zweiten Experiment bereits eingesetzten sechs Items zur Erfassung der Sucherfahrung bestehen (vgl. Anh. A.4). Außerdem werden Studierende aus IT-orientierten Studiengängen wie im zweiten Experiment nur in der Anfangsphase ihres Studiums zur Teilnahme zugelassen.

Von einer zusätzlichen Kontrolle der Motivation der Testpersonen durch einen in Aussicht gestellten Gewinn für die beste Suchleistung wird im dritten Experiment abgesehen, da das Szenario zur Vergabe eines solchen Preises recht komplex und nicht mit einigen Worten erklärt wäre. So müsste bspw. berücksichtigt werden, dass die Hälfte der Teilnehmer davon ausgeht unter schlechteren Untersuchungsbedingungen (mit dem schlechteren System) zu arbeiten. Ansonsten könnte ein Preis für die beste Suchleistung u.U. für diese Treatmentgruppe sogar demotivierend wirken. Stattdessen haben alle Probanden die Möglichkeit, an einer Verlosung von drei Geldpreisen im Wert von 20 €, 30 € und 50 € teilzunehmen.

Da die Testausführung im dritten Experiment erneut für mehrere Testpersonen parallel erfolgen soll, muss, wie im zweiten Experiment, sichergestellt werden, dass sich die Teilnehmer während des Tests nicht gegenseitig beeinflussen können. An dieser Stelle ist bspw. die flexible Handhabung der Aufgabenbearbeitungszeit zu nennen, die wie in Abschnitt 6.3.3 beschrieben dazu führen kann, dass einzelne Untersuchungsteilnehmer ihre Suche vorzeitig beenden, weil sie wahrnehmen, dass andere Probanden den Test bereits abgeschlossen haben. Im Rahmen der Überprüfung möglicher Störeinflüsse muss daher wiederum untersucht werden, ob eine kürzere Bearbeitungszeit einen signifikanten Einfluss auf die Qualitätswahrnehmung der Testkandidaten hat.

In Anlehnung an Al-Maskari und Sanderson (2010) wird bei der Auswertung der Benutzerzufriedenheit, über die Suchdauer hinaus, der Einfluss 23 ausgewählter Benutzerleistungsmaße im Rahmen einer Kovarianzanalyse berücksichtigt. Diese lassen sich in drei Gruppen einteilen, je nachdem ob sich die Leistungsmaße auf die Effektivität, den Aufwand oder die wahrgenommene Relevanz beziehen. Die Gruppe der Effektivitätsmaße umfasst die Zeit, die die Testpersonen benötigen, um das erste richtig relevante Dokument zu finden (S05), die Menge der als relevant bewerteten Dokumente (M10) sowie die Menge der in Übereinstimmung mit den Juroren als relevant bewerteten Dokumente (M16). Der Aufwand der Nutzer hingegen wird durch die Dauer der Suchsitzung (S04), die Anzahl der innerhalb dieser Sitzung durchgeführten Suchen (S01), sowie die erste und letzte betrachtete Rankingposition (S02 u. S03) quantifiziert. Die wahrgenommene Qualität der Suchergebnisse schließlich wird anhand der mittleren Bewertungen relevanter (B04, B05 u. B06) bzw. irrelevanter (B01 u. B03) Dokumente festgestellt. Da die Relevanz im dritten Experiment sowohl anhand einer binären als auch anhand einer 4-stufigen Skala erfasst wird

(vgl. Abschn. 7.3.2), kommen weitere Kovariaten hinzu, welche eine etwas konservativere Betrachtungsweise der Suchleistung ermöglichen. Tabelle D.7 in Anhang D.4 enthält eine Übersicht aller im dritten Experiment zusätzlich verwendeten Kovariaten. An dieser Stelle soll kurz angemerkt werden, dass die Einbeziehung der leistungsbezogenen Kovariaten methodisch nicht ganz unproblematisch sein kann. Eine wesentliche Annahme im Rahmen einer Kovarianzanalyse ist es gerade, dass der Wert der Kovariaten unabhängig von der Treatmentgruppe ist (vgl. Abschn. 4.3.2.3). Diese Unabhängigkeit ist für die Benutzerleistung nicht mehr ohne Weiteres gegeben und muss daher in jedem einzelnen Fall überprüft werden (vgl. Abschn. 7.4.6.2).

Als letzter Punkt soll an dieser Stelle der Umgang mit einem möglichen Einfluss der Aufgabenstellung und damit verbunden der freien Wahl von Suchbegriffen erörtert werden. Wie bei der Beschreibung der ersten beiden Experimente bereits erwähnt, wird die Abfolge der Aufgaben innerhalb der Versuchsgruppen randomisiert, um Lern- und Reihenfolgeeffekte zu kontrollieren. Die Tatsache, dass die Auswertung der Suchanfragen im Anschluss an die Nutzerstudie das Vorhandensein eines Topiceffekts nahe legt, führt dazu, dass im Frühjahr 2014 einige Tests nachgeholt werden müssen (vgl. Abschn. 7.4.1). Der beobachtete Topiceffekt besteht darin, dass einem beträchtlicher Anteil der Testpersonen der Begriff eines Wikis unbekannt zu sein scheint. Dies äußerte sich zum einen durch konkretes Nachfragen einiger Probanden im Testverlauf, zum anderen fällt im Zuge der Auswertung der Suchanfragen auf, dass erstaunlich viele Versuchspersonen bei der Bearbeitung der Suchaufgabe, in der es um den Einsatz von Wikis im Schulunterricht geht (vgl. Abschn. 7.3.4), zunächst nur den Begriff *Wiki* in das Suchfeld eingeben. Um dieser Problematik im Rahmen der Nachtests vorzubeugen, werden Einzeltests durchgeführt und die Teilnehmer erhalten vor jeder Aufgabe die Möglichkeit, Fragen zur Aufgabenstellung zu klären.

Zusammengefasst werden somit im dritten Experiment im Wesentlichen dieselben Kovariatengruppen abgedeckt wie im zweiten Experiment, jedoch mit der Einschränkung, dass einige Variablen aus untersuchungs-ökonomischen Gründen diesmal nicht mit erhoben werden. Um welche Variablen es sich hierbei im Einzelnen handelt ist in Anhang B.10 näher beschrieben. Zusätzlich werden diesmal einige Benutzerleistungskovariaten berücksichtigt, um den Einfluss der eigenen Suchleistung auf die Ergebnisse mit betrachten zu können. Es lässt sich somit festhalten, dass auch im dritten Experiment eine Vielzahl möglicher Störeinflüsse berücksichtigt und unter möglichst standardisierten Bedingungen getestet werden.

7.3.4. Aufbau des Testkorpus

Als Datenbasis für die verwendeten Suchsysteme bedarf es im dritten Experiment erneut eines Testkorpus, das nach Relevanz bewertete Dokumente enthält. Analog zum zweiten Experiment wird in dieser Untersuchung auf Webdokumente zurückgegriffen, um den Testpersonen eine größtmögliche Realitätsnähe zu vermitteln. Im Wesentlichen sind dabei die gleichen Arbeitsschritte durchzuführen, wie in Abschnitt 6.3.4 bereits beschrieben. Deshalb soll an dieser Stelle nur auf einige Besonderheiten des dritten Experiments eingegangen werden.

In Bezug auf die Testaufgaben wird der Vergleichbarkeit halber zunächst geprüft, ob die im zweiten Experiment gestellten Suchaufgaben wiederverwendet werden können. Da jedoch vor dem Hintergrund der tiefgreifenden energiepolitischen Änderungen der letzten Jahre davon ausgegangen werden muss, dass sich im Bereich der Förderung für heizungsunterstützende Solarthermieanlagen in den vier Jahren seit dem zweiten Experiment einige Änderungen erge-

ben haben, wird von einer Wiederverwendung der ersten Aufgabe abgesehen. Anders sieht es dagegen bei der Aufgabe aus, sich über die Geräuschentwicklung von Windkraftanlagen zu informieren. Diese Aufgabe ist weniger auf die Bearbeitung eines speziellen Aspekts (Suche nach einer geeigneten Fördermöglichkeit) als vielmehr auf die Schaffung eines allgemeinen Überblicks gerichtet, sodass die Aktualität der Dokumente bei der zweiten Aufgabe eine geringere Rolle spielt. Um trotzdem die Aktualität der Ergebnislisten sicherzustellen, werden die im zweiten Experiment verwendeten Daten um aktuelle Dokumente ergänzt.

Mit Blick auf die Tatsache, dass das Lehramtsstudium einen wichtigen Studienschwerpunkt der Universität Hildesheim bildet, wird der Schwerpunkt im dritten Experiment darüber hinaus auf lehramtsrelevante Themen gelegt.

Konkret lauten die im dritten Experiment verwendeten Suchaufgaben wie folgt:

1. Stellen Sie sich vor, in der Nähe Ihres Wohnortes soll eine Windkraftanlage gebaut werden. Informieren Sie sich im Detail über die Geräuschentwicklung von Windkraftanlagen.

2. Stellen Sie sich vor, Sie halten ein Überblicksreferat zum Thema „Englischunterricht in der Grundschule“. Informieren Sie sich im Detail über Vor- und Nachteile des Englischunterrichts in der Grundschule.

3. Stellen Sie sich vor, Sie sind Lehrer/in und wollen zum ersten Mal ein Wiki im Unterricht einsetzen. Informieren Sie sich im Detail über Vor- und Nachteile der Nutzung von Wikis im Schulunterricht.

Die Beurteilung der Relevanz der Dokumente erfolgt durch fünfzehn Juroren. Wie in Abschnitt 7.3.1 bereits angedeutet, stehen den Juroren dabei in Ergänzung zu den Bewertungen relevant und irrelevant die Bewertungskategorien eher relevant und eher irrelevant zur Verfügung. In Anlehnung an Sormunen (2000, S. 63) und Sormunen (2002, S. 325) sind die einzelnen Abstufungen folgendermaßen definiert:

relevant: In dem Dokument werden viele Aspekte des Themas umfassend und differenziert dargestellt. Im Falle eines Themas mit unterschiedlichen Facetten werden die meisten dieser Unterthemen besprochen.

eher relevant: Das Dokument enthält mehr Information, als die Aufgabenbeschreibung, aber die Darstellung des Themas bleibt unvollständig. Im Falle eines Themas mit unterschiedlichen Facetten bleibt die Darstellung auf einige wenige Unterthemen beschränkt.

eher irrelevant: Das Dokument verweist lediglich auf das Thema. Es enthält keine Information, die über die Aufgabenbeschreibung hinaus geht.

irrelevant: Das Dokument enthält keinerlei Information zum Thema.

Im Rahmen des Experiments ermöglicht diese 4-stufige Skala eine feinere Abstimmung der Ergebnislistengüte (vgl. Abschn. 7.3.1). Weiterhin eröffnet sie die Möglichkeit, die Kontextabhängigkeit der Relevanzbewertung besser zu analysieren. Die Ergebnisse der ersten beiden Experimente zeigen, dass die Toleranz für weniger relevante Dokumente bei Benutzern des besseren Systems geringer ausfällt als bei Benutzern des schlechteren Systems. Konkret zeigt sich

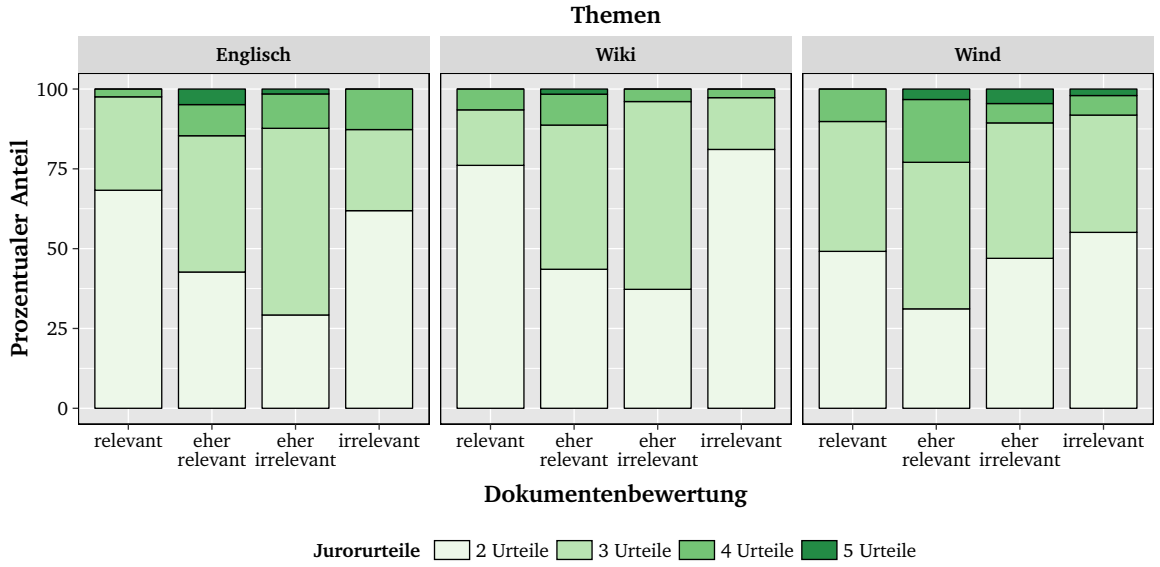


Abb. 7.2.: Prozentuale Verteilung der Jurorenurteile auf die vier Bewertungskategorien. In das finale Korpus fließen nur Dokumente ein, für die maximal drei Jurorenurteile zu Bestimmung der Relevanz erforderlich sind.

dies daran, dass Benutzer des besseren Systems in beiden vorangegangenen Nutzerstudien signifikant mehr relevante Dokumente als irrelevant bewerten (vgl. Abschn. 5.4.2 u. 6.4.3). Im dritten Experiment kann nun direkt überprüft werden, ob Testpersonen, die mit dem besseren System arbeiten, möglicherweise nur noch sehr relevante Dokumente akzeptieren.

Zur Gewährleistung der Objektivität des Testkorpus werden alle Dokumente unabhängig von mindestens zwei Juroren beurteilt, wovon einer stets die Verfasserin der vorliegenden Arbeit ist (im Folgenden als Juror 1 bezeichnet). Bei Nichtübereinstimmung werden weitere Juroren hinzugezogen und die Relevanz schließlich per Mehrheitsentscheid bestimmt. Zusätzlich wird die Reihenfolge der zu bewertenden Dokumente randomisiert, um Lern- und Positionseffekte auszugleichen. Um die Konsistenz innerhalb der Aufgaben zu erhöhen, bewerten die einzelnen Juroren nicht notwendigerweise Dokumente zu allen drei Suchthemen sondern gezielt eine größere Anzahl von Dokumenten zu einer Aufgabe. Während also Juror 1 alle im Korpus enthaltenen Dokumente beurteilt, bewerten die übrigen Juroren im Durchschnitt 46 Dokumente pro Topic.

Tab. 7.1.: Beschreibung des Ausgangskorpus. Für jedes Thema sind die Dokumentenanzahlen aufgeschlüsselt nach relevanten, eher relevanten, eher irrelevanten und irrelevanten Dokumenten angegeben. Des Weiteren ist die mittlere Anzahl benötigter Jurorenurteile pro Dokument vermerkt, um dessen Relevanz zu klassifizieren. Da mindestens zwei übereinstimmende Jurorenurteile verlangt werden, beträgt die minimale Urteilsanzahl 2.

	Wind		Wiki		Englisch	
Urteil	**Anz. Dok.**	**∅ Anz. Urteile**	**Anz. Dok.**	**∅ Anz. Urteile**	**Anz. Dok.**	**∅ Anz. Urteile**
relevant	59	2,61	46	2,30	41	2,34
eher relevant	61	2,95	62	2,69	82	2,77
eher irrelevant	66	2,68	51	2,67	65	2,85
irrelevant	49	2,55	37	2,22	63	2,51
gesamt	235	2,71	196	2,51	251	2,65

Als Bewertungsschema werden die bereits beschriebenen von Sormunen (2002) vorgeschla-

Tab. 7.2.: Interrater-Reliabilität zur Konsistenzprüfung der Suchaufgaben. Angegeben sind sowohl die erreichten Werte für Cohens Kappa als auch die unkorrigierten prozentualen Übereinstimmungswerte. Darüber hinaus werden der Konsistenzprüfung verschiedene Relevanzskalen zugrunde gelegt: 4-stufig, 3-stufig und binär.

		4-stufig		**3-stufig**		**binär**	
	Anz.	**rel. Übereinst.**	**Kappa**	**rel. Übereinst.**	**Kappa**	**rel. Übereinst.**	**Kappa**
				Windthema			
Juror 2	31	0,45	0,24	0,52	0,26	0,74	0,44
Juror 3	46	0,50	0,32	0,67	0,45	0,76	0,50
Juror 4	46	0,63	0,50	0,72	0,55	0,80	0,61
Juror 5	45	0,49	0,29	0,64	0,31	0,73	0,44
Juror 6	48	0,40	0,23	0,52	0,23	0,77	0,55
Juror 7	106	0,42	0,23	0,66	0,41	0,67	0,34
Juror 8	59	0,47	0,27	0,63	0,36	0,75	0,50
Juror 9	20	0,35	0,11	0,50	0,05	0,75	0,48
Mittelwert	50	0,46	0,27	0,61	0,33	0,75	0,48
				Wikithema			
Juror 5	32	0,50	0,31	0,56	0,32	0,91	0,79
Juror 6	24	0,25	0,05	0,33	0,00	0,88	0,69
Juror 7	90	0,47	0,31	0,61	0,39	0,77	0,53
Juror 10	48	0,48	0,33	0,58	0,34	0,69	0,39
Juror 11	40	0,28	0,11	0,35	0,00	0,55	0,25
Juror 12	46	0,46	0,25	0,52	0,26	0,85	0,63
Juror 15	15	0,93	0,91	0,93	0,90	1,00	1,00
Mittelwert	42	0,48	0,32	0,56	0,32	0,80	0,61
				Englischthema			
Juror 3	47	0,38	0,17	0,45	0,12	0,72	0,39
Juror 6	49	0,24	0,01	0,47	0,04	0,73	0,48
Juror 7	56	0,27	0,06	0,43	0,04	0,59	0,18
Juror 8	47	0,66	0,54	0,74	0,61	0,81	0,62
Juror 9	39	0,56	0,41	0,64	0,41	0,90	0,79
Juror 10	21	0,38	0,21	0,52	0,32	0,48	0,05
Juror 12	48	0,44	0,25	0,50	0,18	0,77	0,55
Juror 13	45	0,56	0,39	0,64	0,45	0,80	0,54
Juror 14	63	0,51	0,32	0,63	0,27	0,78	0,55
Mittelwert	46	0,44	0,26	0,56	0,27	0,73	0,46

genen Bewertungskategorien verwendet. Pro Aufgabe umfasst das Korpus zwischen 200 und 250 Dokumente, wovon etwas weniger als die Hälfte der Dokumente auf die beiden extremen Bewertungskategorien entfallen (vgl. Tab. 7.1). Abbildung 7.2 zeigt den prozentualen Anteil der Dokumente mit zwei, drei, vier oder fünf Jurorbewertungen für die vier Bewertungskategorien im gesamten Korpus. Die Beurteilung von Dokumenten der beiden extremen Bewertungskategorien scheint vergleichsweise leichter zu sein, als die Einordnung von Dokumenten in den mittleren Relevanzbereichen. So reichen in den beiden extremen Bewertungskategorien in 60 % der Fälle zwei Relevanzurteile zur Einordnung aus. In über 80 % der Fälle steht jedoch auch im mittleren Relevanzbereich die Entscheidung nach nur drei Relevanzurteilen fest, wobei an dieser Stelle anzumerken ist, dass der Beurteilungsprozess im Schnitt bei 7 % der ursprünglich gesammelten Dokumente (Thema 1: 2 %, Thema 2: 2 %, Thema 3: 17 %) nicht abgeschlossen wird, da schon genügend endgültig bewerte Dokumente zur Fertigstellung des Korpus vorhanden sind.

Zur Überprüfung der Konsistenz des Korpus wird wie im zweiten Experiment (vgl. Abschn. 6.3.4) die Interrater-Reliabilität bestimmt. Als Maß für die Beurteilung des Einflusses individueller Deutungsunterschiede des in dieser Korpuserstellung verwendeten Bewertungsschemas wird

Tab. 7.3.: Konsistenzprüfung der Neubewertung im Kontext des Windthemas.

	Bewertung Experiment 2					
Bewertung	**relevant**		**irrelevant**		**Gesamt**	
Experiment 3	Anz.	%	Anz.	%	Anz.	%
relevant	14	25	1	3	15	16
eher relevant	23	42	3	8	26	28
eher irrelevant	15	27	13	33	28	30
irrelevant	3	5	22	56	25	27

erneut Cohens Kappa verwendet (vgl. Abschn. 4.1.3.3). Da die Überschneidungen der Juroren untereinander in Bezug auf die bewerteten Dokumente ähnlich wie im zweiten Experiment gering ausfallen, wird Cohens Kappa im dritten Experiment wiederum ausschließlich im Vergleich zu den Bewertungen von Juror 1 berechnet. Aus diesem Grund wird hier im Wesentlichen die Konsistenz der übrigen Juroren in Bezug auf die Relevanzurteile von Juror 1 überprüft. Die Werte für die Interrater-Reliabilität der einzelnen Suchthemen können Tabelle 7.2 entnommen werden. Die Berechnung von Cohens Kappa wird dabei auf drei unterschiedliche Arten vorgenommen. Der erste Kappa-Wert bezieht sich auf die vollen vier Stufen des zugrunde gelegten Bewertungsschemas, während zur Berechnung der anderen beiden Werte die ursprünglich 4-stufige Bewertungsskala in eine 3-stufige bzw. binäre Skala umgerechnet wird.

Legt man die 4-stufige Bewertungsskala zugrunde, fällt die resultierende Übereinstimmung eher gering aus. Die durchschnittlichen Kappa-Werte liegen zwischen $0{,}26$ und $0{,}32$. Um zu ermitteln, worauf diese geringe Übereinstimmung in Bezug auf Cohens Kappa zurückzuführen ist, werden als nächstes die beiden alternativen Skalenvarianten in die Betrachtung einbezogen. Fasst man die beiden mittleren Relevanzstufen (eher relevant u. eher irrelevant) zu einer gemeinsamen Kategorie, zeigt sich zunächst keine große Verbesserung im Vergleich zur ersten Skalenvariante. Für die dreistufige Bewertungsskala variieren die berechneten durchschnittlichen Kappa-Werte für die drei Suchthemen zwischen $0{,}27$ und $0{,}33$.

Im nächsten Schritt wird nun die 4-stufige in eine binäre Skala umgerechnet, was, wie Tabelle 7.2 zu entnehmen, zu einem erheblich verbesserten Wert für Cohens Kappa führt. Auf diese Weise wird in allen drei Fällen ein durchschnittlicher Kappa-Wert von über $0{,}45$ erreicht, was einer annehmbaren Übereinstimmung entspricht (Greve und Wentura, 1997, S. 111; Bortz und Döring, 2006, S. 277). Für Thema 3 liegt der Wert sogar bei $0{,}61$, was knapp einer guten Übereinstimmung entspricht. Der unkorrigierte prozentuale Anteil an Übereinstimmungen ist für alle drei Skalenvarianten höher und sollte deshalb wie im Fall des zweiten Experiments ebenfalls in die Betrachtung einbezogen werden. Im Vergleich ergibt sich damit für die dritte Skalenvariante eine ähnlich hohe prozentuale Übereinstimmung wie im zweiten Experiment (zweites Experiment: $72\,\%$ vs. drittes Experiment: $76\,\%$).

Als zusätzliche Qualitätskontrolle werden für das sowohl im zweiten als auch im dritten Experiment verwendete Thema (Windenergie) die endgültigen Relevanzurteile der in beiden Korpora enthaltenen Dokumente verglichen. Tabelle 7.3 zeigt dazu die Verteilung der im zweiten Experiment als relevant bzw. irrelevant bewerteten Dokumente über die vier Bewertungskategorien der aktuellen Nutzerstudie. Die Gesamtanzahl der in beiden Korpora enthaltenen Dokumente beträgt 94, wovon ursprünglich 55 Dokumente ($59\,\%$) als relevant und 39 Dokumente ($41\,\%$) als irrelevant bewertet sind. Im Zuge der Neubewertung werden $27\,\%$ als irrelevant, $30\,\%$ als eher

Tab. 7.4.: Beschreibung des verwendeten Testkorpus. Für jedes Thema sind neben den jeweils erreichten Kappa-Werten die Dokumentenanzahlen aufgeschlüsselt nach relevanten, eher relevanten, eher irrelevanten und irrelevanten Dokumenten angegeben.

		Wind	**Wiki**	**Englisch**
Kappa	4-stufig	0,54	0,80	0,82
	3-stufig	0,59	0,78	0,78
	binär	0,77	1,00	1,00
Anz. Dok.	relevant	31	31	31
	eher relevant	31	31	31
	eher irrelevant	26	26	26
	irrelevant	27	27	27
	gesamt	115	115	115

irrelevant, $28\,\%$ als eher relevant und $16\,\%$ als relevant bewertet. Wird die 4-stufige Skala zu Vergleichszwecken in eine binäre Skala umgewandelt, ergibt sich eine prozentuale Übereinstimmung zwischen den beiden Testkorpora von $77\,\%$. Der berechnete Kappa-Wert beträgt $0{,}54$. In diesem Zusammenhang ist jedoch zu beachten, dass lediglich im zweiten Experiment als irrelevant eingeordnetes Dokument in der aktuellen Jurorbewertung als relevant eingeschätzt wird. Weiterhin ändert sich bei nur drei Dokumenten die Bewertung von irrelevant auf eher relevant. Bei den 18 relevanten Dokumenten die im Gegensatz zum zweiten Experiment nun als irrelevant oder eher irrelevant eingestuft werden kann auch die Aktualität der Dokumente eine Rolle spielen. Vor diesem Hintergrund erscheint also die binäre Relevanzbewertung der Dokumente als relativ stabil in Bezug auf die durchgeführte Neubewertung.

Zusammenfassend zeigen die Ergebnisse zur Interrater-Reliabilität, dass in Bezug auf die binäre Bewertungsskala eine annehmbare Übereinstimmung der Relevanzurteile vorliegt, die sich auf demselben Niveau wie im zweiten Experiment bewegt (Greve und Wentura, 1997, S. 111; Bortz und Döring, 2006, S. 277). Dies ist insbesondere deshalb von Bedeutung, da von dieser binären Bewertung die Konstruktion der Ergebnislisten und damit die Manipulation der Systemleistung abhängt. Die Konsistenz der 4-stufigen Relevanzskala hingegen fällt im Vergleich geringer aus. Wie bereits beschrieben hängt dies vor allem mit der Schwierigkeit zusammen, Dokumente in den mittleren beiden Bewertungskategorien einzuordnen. Im Vergleich zu den extremen Bewertungskategorien liegt der Anteil an Dokumenten für die mehr als zwei Jurorenurteile zur Einschätzung der Relevanz benötigt werden um $20\,\%$ höher. Unter der Annahme, dass Dokumente umso eindeutiger in eine Relevanzkategorie fallen je weniger Jurorenurteile zu ihrer Einordnung notwendig sind, werden in das finale Korpus ausschließlich solche Dokumente aufgenommen, für deren Beurteilung im Höchstfall drei Relevanzurteile benötigt werden. Wie Tabelle 7.4 zu entnehmen, resultiert dies, wie zu erwarten, in erheblich verbesserten Kappa-Werten, für alle drei Suchaufgaben, die sich im Fall der 4-stufigen Skala nun zwischen $0{,}54$ und $0{,}82$ bewegen. Für alle drei Suchthemen sind im finalen Korpus jeweils 31 relevante, 31 eher relevante, 26 eher irrelevante und 27 irrelevante Dokumente enthalten (vgl. Tab. 7.4). Somit ist auch für Experiment 3 die Anzahl der Dokumente in den einzelnen Bewertungskategorien groß genug gewählt um eine Variation der, in den Ergebnislisten enthaltenen, Dokumente zu ermöglichen.

7.3.5. Beschreibung des Testsystems

Im grundlegenden Aufbau und in den bereitgestellten Funktionen bleibt das im dritten Experiment verwendete Testsystem im Vergleich zum zweiten Experiment unverändert. Einzig die Interaktion zwischen Browser und Server wird, aus Gründen der Geschwindigkeit und der Datenverarbeitung, auf eine asynchrone Kommunikation mit Hilfe von Ajax[1] umgestellt. Wie im zweiten Experiment wird auch in der aktuellen Untersuchung für jede neue Suchanfrage eine neue Ergebnisliste generiert. Dabei werden ähnliche Suchanfragen derselben Person erneut mit derselben Ergebnisliste beantwortet. Zusätzlich wird das Design der Suchseite an die aktuellen Sehgewohnheiten der Nutzer angepasst und das zuvor verwendete Frameset-Layout durch ein CSS-Layout ersetzt. Die neue Benutzeroberfläche mit angezeigter Suchergebnisliste ist in Abbildung 7.3 zu sehen. Die für die Beschreibung der Dokumente erforderlichen Elemente wie Titel, Snippet und URL können wie im Fall des zweiten Experiments während der Korpuserstellung von Google übernommen werden. Da während der Pretestphase erneut zu beobachten ist (vgl. Abschn. 7.3.7), dass einige der mit WinHTTrack gespeicherten Webseiten mit Frame-Escape-Skripten arbeiten, erfolgt die Darstellung der Webseiten in diesem Experiment standardmäßig in einem neuen Tab des Browsers. Dabei wird der Browser so konfiguriert, dass ein versehentliches Schließen von Testsystem und Browser nicht möglich ist. Um darüber hinaus die Tatsache zu verschleiern, dass die angezeigten Webseiten von demselben Server ausgeliefert werden auf dem das Testsystem gehostet ist, wird weiterhin die Adressleiste des Browsers ausgeblendet. Die Relevanzbewertung der angesehenen Dokumente erfolgt nachdem das Dokumententab geschlossen wird. Dazu bleibt die Ergebnisliste ausgegraut, bis die Bewertung durch den Benutzer abgeschlossen ist (vgl. Abb. 7.4). Eine weitere Neuerung stellt die direkte Integration der Fragebögen in das Testsystem dar, die mit LimeSurvey[2], einem Open Source Programm zur Erstellung von Online-Umfragen, realisiert wird. Dies ermöglicht es, die Fragebögen im Anschluss an jede Suchaufgabe direkt im Browser anzuzeigen sowie die Logdaten in einer gemeinsamen Datenbank zu bündeln. Eine Beispielseite des Fragebogens ist in Abbildung 7.5 dargestellt. Um sicherzustellen, dass für dieses Experiment die Voraussetzungen einer Doppelblindstudie erfüllt sind, erfolgt die Zuteilung der Probanden zu den Versuchsgruppen zufällig bei Aufruf der Startseite des Testsystems und ist somit weder Teilnehmer noch Testleiter bekannt.

[1] Ajax steht für den englischen Begriff *Asynchronous JavaScript and XML* mit dem Techniken zur asynchronen Datenübertragungen zwischen Server und Webklienten bezeichnet werden, die den Austausch von Daten ohne das vollständige Neuladen einer Webseite ermöglichen (Wenz, 2007).

[2] `https://www.limesurvey.org/de/`

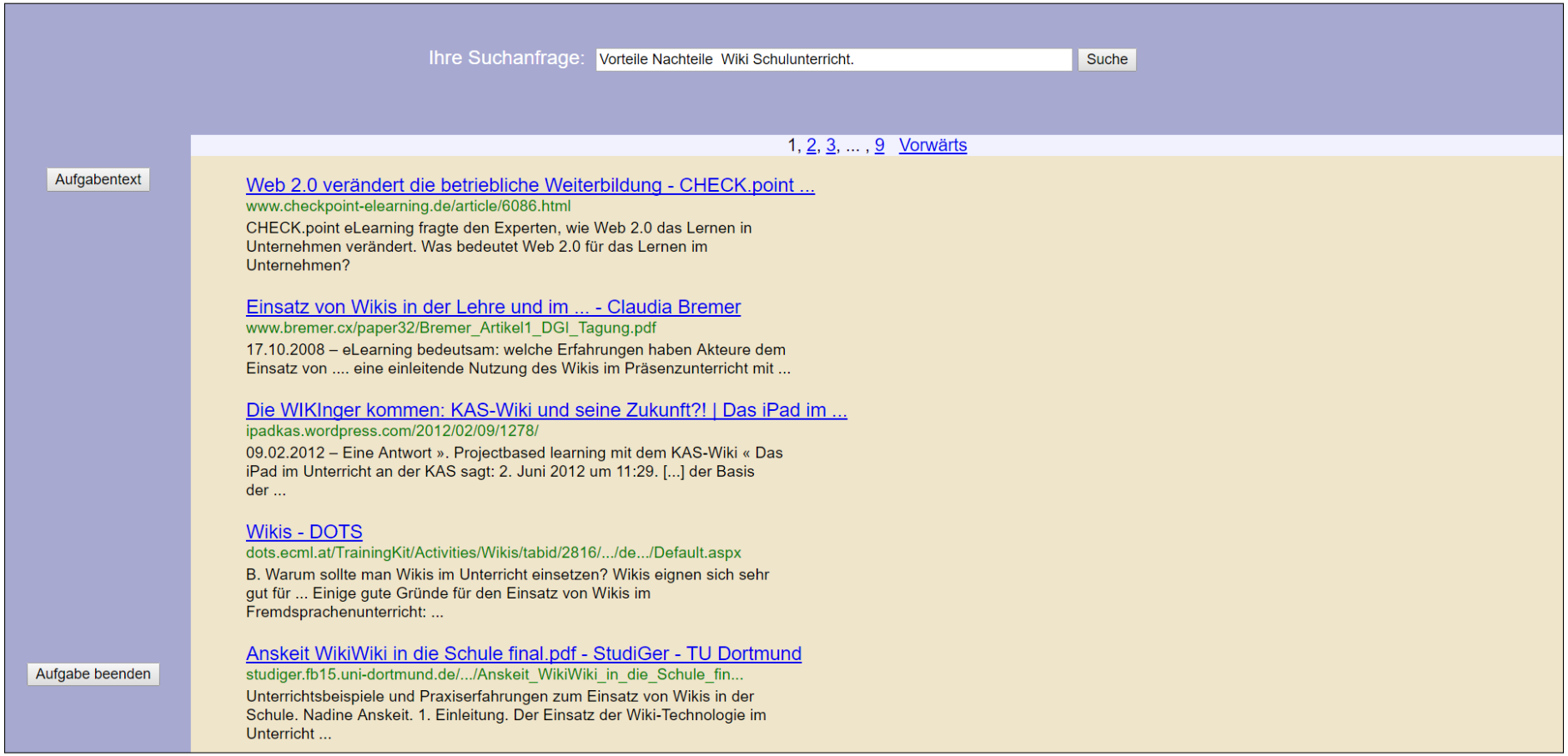


Abb. 7.3.: Testsystem des dritten Experiments: Darstellung der Suchergebnisliste.

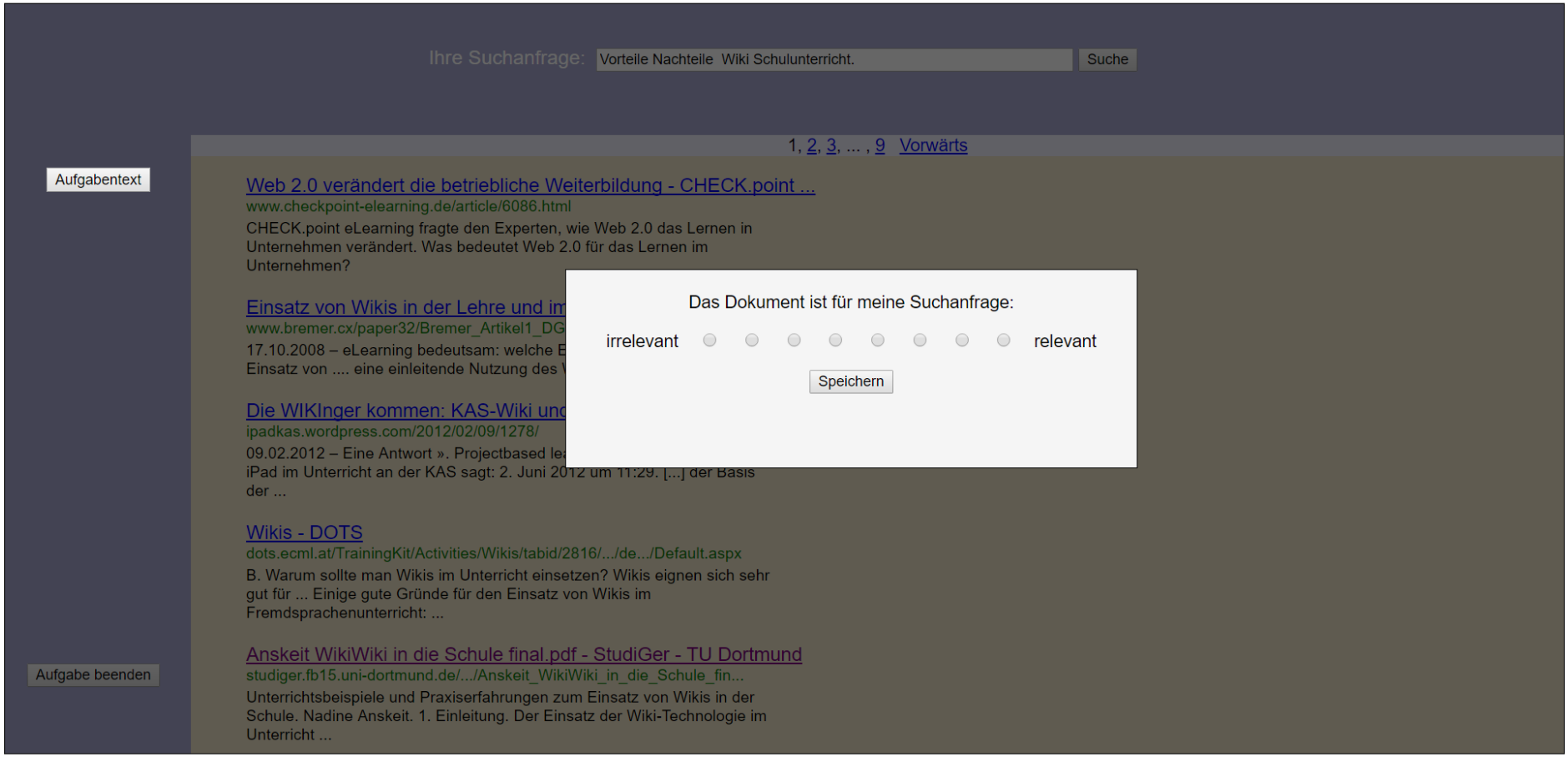
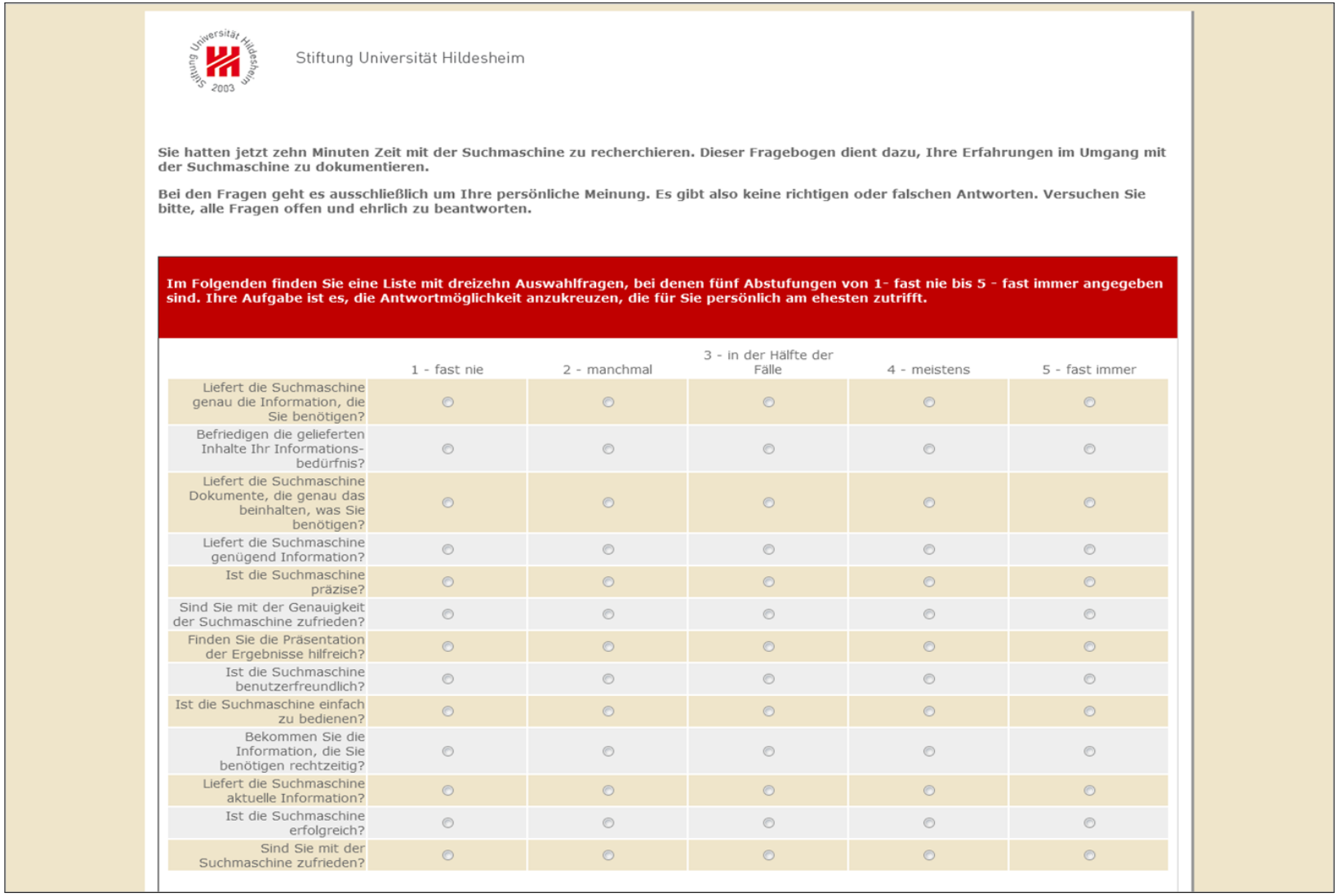


Abb. 7.4.: Testsystem des dritten Experiments: Relevanzbewertung.



Stiftung Universität Hildesheim

Sie hatten jetzt zehn Minuten Zeit mit der Suchmaschine zu recherchieren. Dieser Fragebogen dient dazu, Ihre Erfahrungen im Umgang mit der Suchmaschine zu dokumentieren.

Bei den Fragen geht es ausschließlich um Ihre persönliche Meinung. Es gibt also keine richtigen oder falschen Antworten. Versuchen Sie bitte, alle Fragen offen und ehrlich zu beantworten.

Im Folgenden finden Sie eine Liste mit dreizehn Auswahlfragen, bei denen fünf Abstufungen von 1- fast nie bis 5 - fast immer angegeben sind. Ihre Aufgabe ist es, die Antwortmöglichkeit anzukreuzen, die für Sie persönlich am ehesten zutrifft.

	1 - fast nie	2 - manchmal	3 - in der Hälfte der Fälle	4 - meistens	5 - fast immer
Liefert die Suchmaschine genau die Information, die Sie benötigen?	○	○	○	○	○
Befriedigen die gelieferten Inhalte Ihr Informations-bedürfnis?	○	○	○	○	○
Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	○	○	○	○	○
Liefert die Suchmaschine genügend Information?	○	○	○	○	○
Ist die Suchmaschine präzise?	○	○	○	○	○
Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	○	○	○	○	○
Finden Sie die Präsentation der Ergebnisse hilfreich?	○	○	○	○	○
Ist die Suchmaschine benutzerfreundlich?	○	○	○	○	○
Ist die Suchmaschine einfach zu bedienen?	○	○	○	○	○
Bekommen Sie die Information, die Sie benötigen rechtzeitig?	○	○	○	○	○
Liefert die Suchmaschine aktuelle Information?	○	○	○	○	○
Ist die Suchmaschine erfolgreich?	○	○	○	○	○
Sind Sie mit der Suchmaschine zufrieden?	○	○	○	○	○

Abb. 7.5.: Beispielseite des im dritten Experiment verwendeten Online-Fragebogens.

7.3.6. Ablauf

Ebenso wie das Untersuchungsdesign eine Weiterentwicklung der beiden vorangegangenen Nutzerstudien darstellt, enthält der Ablauf der dritten Untersuchung ebenfalls Elemente der beiden vorherigen Experimente (vgl. Abb. 7.6). Erneut werden die Untersuchungsteilnehmer nicht individuell im Einzeltestverfahren, sondern wie im zweiten Experiment mehrere Teilnehmer parallel im gleichen Raum getestet. Damit der Ablauf für alle Testkandidaten identisch ist, erfolgt die Kommunikation mit den Probanden großenteils standardisiert durch das Testsystem, welches die Teilnehmer mit Hilfe von Instruktionstexten und -videos durch das Experiment führt. Weitergehende Verständnisfragen dürfen jedoch wie in den ersten beiden Experimenten gestellt werden. Im Anschluss an eine audiovisuellen Testinstruktion, die zu Beginn über ein Einführungsvideo erfolgt, werden die Testpersonen gebeten nacheinander drei Suchaufgaben in unterschiedlicher Reihenfolge zu bearbeiten. Je nach Zuordnung zu der Versuchsgruppe erhalten die Probanden, nach Eingabe ihrer individuellen Suchanfrage eine Trefferliste in der entsprechenden Systemgüte. Wie in den vorangegangenen beiden Experimenten steht es den Versuchspersonen nun frei, so viele Dokumente anzuschauen und zu bewerten, wie sie wünschen. Formulieren sie ihre Suchanfrage um, wird eine neue zufällig erzeugte Liste gleicher Systemgüte angezeigt. Die maximale Bearbeitungszeit für jede Aufgabe beträgt dabei erneut zehn Minuten. Nach Ablauf dieser Zeit werden die Untersuchungsteilnehmer automatisch zu dem in Abschnitt 7.3.2 beschriebenen Fragebogen zur Erfassung der Benutzerzufriedenheit weitergeleitet. Bevor die Probanden jedoch mit der Bearbeitung der nächsten Suchaufgabe beginnen, werden die Teilnehmer zunächst gebeten, ihre Erwartungen in Bezug auf die Bearbeitung dieser Aufgabe darzulegen (vgl. Abschn. 7.3.1). Um auch auf die aus der Bearbeitung der dritten Aufgabe resultierende Erwartungshaltung zugreifen zu können, wird in diesem Fall nach den Erwartungen bezüglich einer Weiterarbeit mit der betreffenden Suchmaschine gefragt. Notwendige Angaben zur Person werden am Ende des Experiments über einen zusätzlichen Fragebogen erhoben. In der letzten offenen Frage wird wie in den vorangegangenen beiden Experimenten nach weiteren Anmerkungen der Untersuchungsteilnehmer gefragt. Als Anreiz zur Experimentteilnahme, werden unter allen Versuchspersonen drei Geldpreise verlost. Die Gesamtdauer der Untersuchung beträgt ca. 45 Minuten.

7.3.7. Ergebnisse des Pretests

Im Pretest des dritten Experiments wird das gesamte experimentelle Material auf mehreren Ebenen getestet. Der Einsatz der Methode des lauten Denkens soll hier mögliche Fehlerquellen bei der Bedienung des Testsystems aufzeigen und helfen, Verständnisprobleme hinsichtlich der Instruktionstexte und Fragebögen zu identifizieren. Ergänzend werden die Versuchspersonen wie zuvor bei der Bearbeitung der Aufgaben beobachtet. Im Gegensatz zu den ersten beiden Experimenten finden die Überprüfung des Untersuchungsmaterials und die Überprüfung des Untersuchungsdesigns jedoch getrennt statt. Damit gliedert sich der Pretest ähnlich wie im zweiten Experiment in zwei Phasen. Während das Ziel der ersten Phase darin besteht, die sprachliche und inhaltliche Qualität der Instruktionstexte und Fragebögen sowie die Funktionalität des Testsystems zu überprüfen, besteht das Ziel der zweiten Phase in erster Linie darin, den Erfolg der Erwartungsmanipulation einschätzen zu können.

Die vier Versuchspersonen der ersten Pretestphase sind wie im Fall des zweiten Experiments Mitarbeiter des Instituts für Informationswissenschaft und Sprachtechnologie der Universität

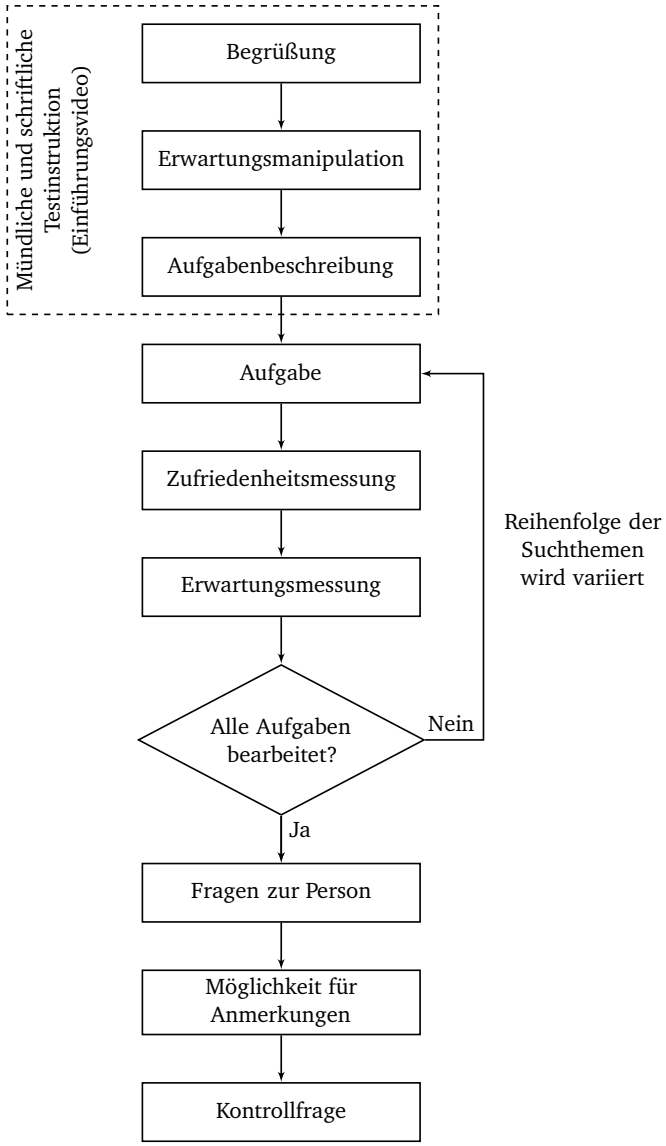


Abb. 7.6.: Schematische Darstellung des Versuchsablaufs des dritten Experiments.

Hildesheim. Sie sind mit der Testpopulation nicht direkt vergleichbar, da davon auszugehen ist, dass sie sowohl im IR-Bereich als auch in Bezug auf die verwendeten Forschungsmethoden über ein breiteres Wissen verfügen als die tatsächlichen Probanden. Trotzdem zeigen die Ergebnisse der ersten Pretestphase, welche Sachverhalte angepasst werden müssen, und führen zu einigen Änderungen bei der Bedienung des Testsystems, wovon die Wesentlichen im Folgenden kurz erläutert werden. Bis auf wenige Ausnahmen, die die in Abschnitt 6.3.7 bereits erläuterten Nachteile des EUCS-Instruments betreffen, werden die Fragebögen von den Teilnehmenden als verständlich und leicht nachvollziehbar bezeichnet, sodass hier keine Änderungen nötig sind. Die ursprüngliche Umsetzung des Testsystems, bei der die Webdokumente wie im zweiten Experiment direkt im Inhaltsbereich des Testsystems angezeigt werden, stellt sich im Verlauf der ersten Pretestphase als unzureichend heraus, da einige der mit WinHTTrack gespeicherten Webseiten ähnlich wie im Fall des zweiten Experiments mit Frame-Escape-Skripten arbeiten. Um die damit verbundenen technischen Schwierigkeiten zu umgehen, werden die Webseiten in der Hauptuntersuchung standardmäßig in einem neuen Tab des Browsers geöffnet. Eine weitere

nennenswerte Änderung ergibt sich aus der Tatsache, dass das gesamte Untersuchungsmaterial des dritten Experiments in digitaler Form vorgehalten wird. Während den Testpersonen die aktuelle Aufgabenbeschreibung in den ersten beiden Experimenten in Papierform vorliegt und jederzeit nachgelesen werden kann, wird sie im Pretest zunächst nur zu Beginn jeder neuen Aufgabe eingeblendet. Im Pretest stellt sich jedoch heraus, dass die Versuchspersonen während der gesamten Bearbeitungszeit auf die Aufgabenbeschreibung zugreifen möchten, sodass der Benutzeroberfläche für die Hauptuntersuchung ein entsprechender Button hinzugefügt wird. In Bezug auf die neue 8-stufige Skala zur Relevanzbewertung zeigt sich, dass die Versuchspersonen im Rahmen der qualitativen Voruntersuchung gut mit dieser erweiterten Skala zurecht kommen, weshalb sie in der Hauptuntersuchung beibehalten wird.

Eine wesentliche Teilnahmevoraussetzung für Versuchspersonen der zweiten Pretestphase ist, dass ihnen die Hintergründe der geplanten Untersuchung unbekannt sind, da die Hauptaufgabe dieser Phase darin besteht, die Qualität der Erwartungsmanipulation zu überprüfen. Insgesamt nehmen neun Versuchspersonen in dieser Phase an dem Pretest teil. Die Ergebnisse der ersten Hälfte der Tests erwecken zunächst den Anschein, dass die Manipulation der Erwartungshaltung zu offensichtlich ist. So geben alle Versuchspersonen im Nachgespräch an, dass sie den Eindruck gehabt hätten, dass die Information darüber, welchem der beiden Systeme sie zugeteilt werden der Beeinflussung dienen soll. Da zu diesem Zeitpunkt nicht bekannt ist, ob tatsächlich die Manipulation zu offensichtlich ist oder die Vorerfahrungen der Versuchspersonen (Doktoranden der Nachbarinstitute) mit psychologischen Studien dafür verantwortlich sind, wird diese Pretestphase noch einmal mit Probanden mit geringerer Testerfahrung wiederholt. Zusätzlich werden die Testinstruktionen jedoch so abgeändert, dass die Erwartungsmanipulation weniger stark in den Fokus der Testpersonen rückt. Dies betrifft zum einen die verwendete Powerpointfolie des Einführungsvideos, bei der auf eine Untergliederung des Textes verzichtet wird, sodass die Manipulation visuell weniger in Erscheinung tritt. Zum anderen wird am Ende der Folie ein weiterer Satz hinzugefügt, der nicht mit der Erwartungsmanipulation im Zusammenhang steht. Um die Erwartungsmanipulation zusätzlich auch auf textueller Ebene abzuschwächen wird anstelle der ursprünglichen Formulierung (*Aktuelle Studien zur Suchmaschinennutzung haben gezeigt, dass Benutzer auch mit unterschiedlich guten Suchmaschinen ähnliche Ergebnisse erreichen.*) die folgende Formulierung gewählt: *Aktuelle Studien zur Suchmaschinennutzung haben gezeigt, dass die Suchmaschinenleistung keinen großen Einfluss auf den Sucherfolg der Benutzer hat. Vielmehr erreichen Benutzer auch mit unterschiedlich guten Suchmaschinen vergleichbare Ergebnisse.* Nach Umsetzung dieser Änderungen und dem Rückgriff auf Probanden mit geringer bis keiner Testerfahrung verläuft die zweite Hälfte des Tests bezüglich der Erwartungsmanipulation erfolgreich. Dies zeigt sich zum einen in einer richtigen Beantwortung der Kontrollfrage und zum anderen an der Tatsache, dass keine der Versuchspersonen im Rahmen des Nachgesprächs zum Ausdruck bringt, dass sie die Information über die Gruppenzuteilung als Beeinflussung wahrgenommen hat.

7.4. Ergebnisse

Der nun folgende Abschnitt stellt die Ergebnisse der Hauptuntersuchung der dritten Nutzerstudie dar. Die einzelnen Benutzertests finden im Herbst 2013 an der Universität Hildesheim statt. Die

in Abschnitt 7.3.3 erwähnte Nacherhebung wird hingegen im Frühjahr 2014 durchgeführt. Im Rahmen der Darstellung der Ergebnisse wird zunächst auf die Zusammensetzung der untersuchten Stichprobe eingegangen. Wie schon in Kapitel 6 folgt anschließend eine kurze Erklärung der relevanten Stichproben und Qualitätsstufen, die bei der Auswertung des dritten Experiments eine Rolle spielen. Im Anschluss daran werden die Ergebnisse in Bezug auf die Benutzerleistung erläutert. Analog zum Vorgehen im Rahmen der zweiten Untersuchung ist auch beim dritten Experiment der eigentlichen Auswertung der Benutzerzufriedenheit eine Faktorenanalyse vorangestellt. Diese dient erneut dazu, die Vielzahl der im Experiment erhobenen Zufriedenheitsitems auf einige wenige zentrale Zufriedenheitsdimensionen zu verdichten. Nach der varianzanalytischen Auswertung wird die Reliabilität der Ergebnisse anhand verschiedener Gütekriterien überprüft, um die Erklärungsbeiträge der zusätzlich wirksamen Einflussfaktoren einschätzen zu können.

7.4.1. Beschreibung der Stichprobe

Die Probanden des dritten Experiments rekrutieren sich nahezu ausschließlich aus Studierenden der Universität Hildesheim. Ein Teil der Tests kann erneut im Rahmen einer regulären Lehrveranstaltung im Bereich der Informationswissenschaft durchgeführt werden. Die übrigen Teilnehmer werden über Mailinglisten, Aushänge und Direktansprache rekrutiert. Insgesamt nehmen 153 Personen an der Untersuchung teil.

Tab. 7.5.: Verteilung der Testteilnehmer auf die Untersuchungsgruppen ($n = 153$). Dargestellt ist die Gruppenverteilung in Abhängigkeit von den im Text beschriebenen Gütekategorien sowie dem Geschlecht (m/w) der Teilnehmer.

		Stichprobe	**System**					
			gut			**schlecht**		
			m	**w**	**ges.**	**m**	**w**	**ges.**
Erwartung	**hoch**	SP_B	3	17	20	8	15	23
		SP_N	2	6	8	4	1	5
		SP_{UV}	2	7	9	1	8	9
		SP_{SB}	1	6	7	1	6	7
		SP_{MV}	1	2	3	0	2	2
		SP_{TD}	0	0	0	0	1	1
	niedrig	SP_B	5	15	20	2	21	23
		SP_N	1	1	2	3	4	10
		SP_{UV}	2	10	12	3	9	12
		SP_{SB}	1	8	9	2	8	10
		SP_{MV}	1	1	2	0	1	1
		SP_{TD}	0	2	2	1	1	2

Auch im dritten Experiment müssen einzelne Fälle nachträglich von der Auswertung ausgeschlossen werden. Im Folgenden werden daher zunächst die zugrunde liegenden Ausschlusskriterien erläutert. Um besser entscheiden zu können, welche Fälle in die Stichprobe aufgenommen werden sollen, erfolgt analog zum zweiten Experiment eine Einteilung der Stichprobe in drei Unterkategorien, je nachdem ob die Daten ohne Einschränkung (Teilstichprobe SP_B), unter Vorbehalt (Teilstichprobe SP_{UV}) oder nicht (Teilstichprobe SP_N) in die Auswertung eingehen können. Tabelle 7.5 gibt die Verteilung der Probanden diese drei Unterstichproben unter zusätzlicher Berücksichtigung des Geschlechts der Untersuchungsteilnehmer wieder. Die erste Unterstich-

probe umfasst alle Fälle, die ohne Einschränkungen in die Untersuchung einbezogen werden können (SP_B). Bei diesen Fällen sind alle drei Aufgaben bearbeitet und die Kontrollfrage ist richtig beantwortet. Dies trifft auf $56{,}2\,\%$ der Gesamtstichprobe zu. Mindestens zwanzig Fälle pro Versuchsgruppe sind somit ohne Einschränkungen wertbar.

Die zweite Unterstichprobe beinhaltet alle Fälle, die von der Untersuchung ausgeschlossen werden müssen (SP_N). Als Ausschlusskriterien gelten technische Probleme, die bspw. dazu führen, dass nicht alle drei Aufgaben bearbeitet werden können (5 Fälle), die Eingabe von Suchbegriffen, die nicht zum Thema passen (15 Fälle) sowie Fälle, in denen nur ein einziges Dokument pro Aufgabe betrachtet wird (5 Fälle). In letzterem Fall wird, wie im zweiten Experiment, unterstellt, dass in diesen Fällen nicht von einer ernsthaften Teilnahme ausgegangen werden kann. Diese Annahme wird zudem durch die Tatsache gestützt, dass in drei der fünf Fälle, auf die dieses Ausschlusskriterium zutrifft, weitere einschränkende Faktoren hinzukommen. An dieser Stelle ist außerdem darauf hinzuweisen, dass die auftretenden technischen Probleme in drei der soeben erwähnten Fälle dazu führen, dass es nicht möglich ist, die demographischen Daten der jeweiligen Testpersonen zu erfassen. Dies erklärt, weshalb die Gesamtzahl der ausgeschlossenen Testpersonen in der letzten Gruppe nicht mit den nach Geschlecht aufgeschlüsselten Anzahlen übereinstimmt (vgl. Tab. 7.5). Insgesamt müssen somit $16{,}3\,\%$ der durchgeführten Test nachträglich von der Auswertung ausgeschlossen werden.

Die dritte Unterstichprobe fasst erneut all diejenigen Fälle zusammen, bei denen nicht auf den ersten Blick erkennbar ist, ob die beobachteten Auffälligkeiten einen verzerrenden Einfluss auf die Untersuchungsergebnisse haben (SP_{UV}). Insgesamt werden $27{,}4\,\%$ der erhobenen Fälle in diese Unterstichprobe eingeordnet. Teilnehmer in dieser Gruppe geben entweder Suchbegriffe ein, die die gestellte Aufgabe nur z.T. abdecken (SP_{SB}), beantworten die Kontrollfrage nicht korrekt (SP_{MV}) oder die Beantwortung der offenen Frage bzw. ein informelles Gespräch im Anschluss an das Experiment deuten darauf hin, dass der Versuchsaufbau durchschaut wird (SP_{TD}). Bis auf vier Fälle, in denen sowohl unspezifische Suchanfragen gestellt als auch die Kontrollfrage nicht richtig beantwortet wird, lassen sich die übrigen Fälle eindeutig einer der drei Kategorien zuordnen. Bei den folgenden Angaben ist daher zu beachten, dass diese vier Fälle doppelt in die Beschreibung der dritten Unterstichprobe eingehen. Da der einschränkende Charakter dieser Unterstichprobe eng mit der Frage nach den Erwartungen der Testpersonen verknüpft ist, soll an dieser Stelle noch etwas weiter ausgeführt werden, welche Bedeutung die jeweiligen Entscheidungen für die weitere Auswertung des Experiments haben. Die meisten der als unter Vorbehalt wertbar klassifizierten Fälle betreffen Testpersonen, deren eingegebene Suchbegriffe die Anforderungen der Aufgabenstellung nur z.T. erfüllen (SP_{SB}: 33 Fälle). Am häufigsten tritt dieses Problem im Zusammenhang mit Thema 3 auf (Stellen Sie sich vor, Sie sind Lehrer/in und wollen zum ersten Mal ein Wiki im Unterricht einsetzen. Informieren Sie sich im Detail über Vor- und Nachteile der Nutzung von Wikis im Schulunterricht.). Die diesbezüglich eingegebenen Suchbegriffe lauten z.B.: *wikis, nutzung wikis, vor- und nachteile nutzung wikis, nutzung von wikis pro contra, nachteile von wikis*. Problematisch sind derartige Sucheingaben aus zwei Gründen: Zum einen können solche Eingaben derart gedeutet werden, dass einige Probanden die Aufgabenbeschreibung nicht gründlich gelesen haben und deshalb die Suche mit dem Teilaspekt beginnen, der Ihnen am ehesten in Erinnerung geblieben ist. Für diese Auslegung spricht neben der Tatsache, dass in den

meisten Fällen bereits die erste gestellte Suchanfrage davon betroffen ist, auch der Umstand, dass einige der fraglichen Suchsessions während der Aufgabenbearbeitung nicht über diese allgemeineren Suchanfragen hinausgehen. Aufgrund der experimentellen Situation ist hierbei zu beachten, dass das Testsystem inhaltlich nicht zwischen den gestellten Suchanfragen unterscheidet, sondern im Falle einer Umformulierung lediglich eine umsortierte Liste der im Testkorpus enthaltenen Dokumente angezeigt wird. Unklar ist nun, welche Erwartungen Testpersonen, die Suchbegriffe wie *wikis* oder *nutzung wikis* verwenden, an die zurückgelieferten Suchergebnisse haben. Nicht in jedem Fall lässt sich schließen, dass Suchergebnisse die alle Aspekte einer Aufgabe abdecken, obwohl nur nach einem Teilaspekt gesucht wird, Misstrauen bei den Probanden in Bezug auf das Untersuchungsdesign auslösen. Zwar mag dies für einige Probanden der Fall sein, weswegen eine Aufdeckung des Versuchsaufbaus im Einzelfall nicht ausgeschlossen werden kann, die alleinige Eingabe unspezifischer Suchbegriffe stellt jedoch noch kein hinreichendes Indiz für diese Interpretation dar. Insbesondere könnten einige Probanden gar nicht bemerkt haben, dass die von ihnen eingegebenen Suchbegriffe die Aufgabenstellung nur z.T. abdecken und daher mit den Suchergebnisse zufrieden sein. Für eine solche Interpretation sprechen z.B. diejenigen Fälle, in denen nur eine einzige Suche durchgeführt wird.

Interessant ist weiterhin die Frage, warum dieses Problem am häufigsten im Zusammenhang mit Thema 3 auftritt. Durch die Variation der Aufgabenreihenfolge (vgl. Abschn. 7.3.3) können Ermüdungseffekte als Grund ausgeschlossen werden. Eine mögliche Interpretation wäre, dass einigen Versuchspersonen der Begriff eines Wikis unbekannt ist und sie sich deshalb zunächst allgemein darüber informieren möchten. Dies stimmt mit der Beobachtung überein, dass in den meisten Fällen bereits die erste gestellte Suchanfrage davon betroffen ist. Zwar ist ein Vorgehen vom Allgemeinen zum Speziellen für den Suchprozess zunächst nicht unüblich, allerdings wird die These zum Bekanntheitsgrad des Begriffs Wiki dadurch gestützt, dass einige Probanden explizit nachfragen was ein Wiki sei. Damit ist es durchaus möglich, dass Teilnehmer, die zum Zeitpunkt des Tests noch nicht über ein entsprechendes Wissen verfügen, den Eindruck bekommen, dass Wikis ausschließlich im Lehrkontext verwendet werden. Folglich kann auch in diesem Fall aus einer unspezifischen Suche nicht direkt auf ein Durchschauen des Versuchsaufbaus geschlossen werden.

Zusammenfassend lässt sich sagen, dass wahrscheinlich beide beschriebenen Szenarien in dem Experiment auftreten. Insbesondere bei Thema 3 deuten jedoch die expliziten Nachfragen der Testpersonen darauf hin, dass der Hauptgrund unspezifischer Suchen im Bekanntheitsgrad des Begriffs Wiki zu sehen ist. Da jedoch in beiden Fällen nicht zwingend davon auszugehen ist, dass das Untersuchungsdesign durchschaut wird und es vielmehr plausibel erscheint, dass eine unspezifische Suche von den Testpersonen nicht bemerkt wird, werden diese Fälle nicht von vornherein von der Auswertung ausgeschlossen. Stattdessen wird untersucht, wie die Hinzunahme dieser Fälle die Ergebnisse beeinflusst. Die in Zusammenhang mit Thema 3 aufgetretenen Schwierigkeiten deuten schon auf einen möglichen Topiceffekt in den Daten hin. Zusätzlich lässt sich nicht ausschließen, dass weitere Teilnehmer zwar unauffällige Suchanfragen stellen, ihnen der Wiki-Begriff aber trotzdem nicht vertraut ist. Aufgrund der Randomisierung der Testaufgaben und der damit einhergehenden Beeinflussung der Erwartungshaltung der Teilnehmer bei den Folgeaufgaben durch Thema 3, ist ein kompletter Ausschluss dieser Aufgabe jedoch

nicht möglich. Aus diesem Grund wird bei der Auswertung wie im zweiten Experiment auf eine Ausgeglichenheit der Versuchsgruppen sowohl hinsichtlich des Suchthemas als auch der Bearbeitungsreihenfolge geachtet (vgl. Abschn. 7.4.2). Um in diesem Zusammenhang eine minimale ausgeglichene Gruppengröße von 20 ohne Einschränkungen wertbaren Fällen sicherstellen zu können, werden im Frühjahr 2014 noch einmal zehn zusätzliche Benutzertests durchgeführt, die bereits in der Darstellung in Tabelle 7.5 mit enthalten sind.

Die anderen beiden Gruppen, bei denen nicht von vornherein zu erkennen ist, ob die beobachteten Schwierigkeiten tatsächlich einen verzerrenden Einfluss auf die Ergebnisse ausüben, sind jeweils mit nur wenigen Fällen vertreten. In insgesamt acht Fällen wird die Kontrollfrage nicht richtig beantwortet, mit der überprüft wird, ob die Versuchspersonen sich an die Erwartungsmanipulation erinnern (SP_{MV}) (vgl. Abschn. 7.3.1). Da eine unbewusste Verarbeitung der Erwartungsmanipulation auch im dritten Experiment nicht ausgeschlossen werden kann, verbleiben diese Fälle wie schon im zweiten Experiment vorerst in der Stichprobe, bis abschließend geklärt ist, ob ihre Hinzunahme einen Einfluss auf die Ergebnisse hat. Bei weiteren fünf Versuchspersonen deuten die Beantwortung der offenen Frage oder ein informelles Gespräch im Anschluss an das Experiment darauf hin, dass der Versuchsaufbau möglicherweise durchschaut wird (SP_{TD}). Auch hier wird zunächst überprüft, ob und wie die Hinzunahme dieser Fälle das Ergebnis beeinflusst, bevor sie vorschnell aus der Stichprobe entfernt werden. Wenn alle Probanden der Teilstichproben SP_B und SP_{UV} in die Auswertung einbezogen werden, wird mit mindestens 29 Testpersonen pro Versuchsgruppe die ursprünglich angestrebte Stichprobengröße leicht übertroffen. Analog zu Experiment 2 wird diese weniger kontrollierte Gesamtstichprobe als SP_A bezeichnet. Weiterhin lässt sich das Geschlecht der Teilnehmer betreffend kein offensichtliches Ungleichgewicht in Bezug auf die Zugehörigkeit zu einer der problematischen Fallgruppen erkennen. Die genaue Verteilung der Probanden auf die vier Versuchsgruppen kann Tabelle 7.5 entnommen werden.

In Tabelle 7.6 sind die demographischen Merkmale der Gesamtstichprobe SP_A zusammengefasst. Die Basis für die empirische Untersuchung bildet im dritten Experiment somit eine Stichprobe von 128 Testpersonen im Alter von 18 bis 34 Jahren. Anzumerken ist an dieser Stelle, dass bei einer anderenfalls ohne Einschränkungen auswertbaren Probandin nach Beendigung aller Testaufgaben ein technisches Problem auftritt, welches dazu führt, dass der Fragebogen zum persönlichen Hintergrund unbeantwortet bleibt. Bis auf das Geschlecht beziehen sich daher alle im Folgenden dargestellten Angaben zur Stichprobenbeschreibung auf die übrigen 127 Testpersonen in der Stichprobe.

Das Medianalter beträgt 21 Jahre. $20{,}3\,\%$ der Probanden sind männlichen Geschlechts im Vergleich zu $79{,}7\,\%$ weiblichen Testpersonen. Bezogen auf diese beiden Kriterien erweist sich die Stichprobe im Vergleich zum zweiten Experiment damit als weitgehend übereinstimmend. Der Anteil der Probanden nichtdeutscher Muttersprache ist etwas höher als in den ersten beiden Experimenten. Auf die Frage, ob Deutsch ihre Muttersprache ist, antworten $81{,}9\,\%$ der Probanden mit ja, $4{,}7\,\%$ geben an zweisprachig aufgewachsen zu sein und $13{,}4\,\%$ im Vergleich zu $13{,}5\,\%$ im ersten und $11{,}5\,\%$ im zweiten Experiment. Dieser Unterschied erscheint jedoch nicht dramatisch und kann als zufällige Schwankung interpretiert werden. Zusätzlich wird ein möglicher Einfluss der Muttersprache auch im dritten Experiment statistisch überprüft (vgl. Abschn. 7.4.6). Mit Ausnahme von zwei Fällen handelt es sich bei den Probanden um Studierende. Wie sich die stu-

Tab. 7.6.: Demographische Daten. Nach der Datenbereinigung liegt eine Stichprobe von $n = 127$ Untersuchungsteilnehmern vor, was einer Stichprobenausschöpfung von $83\,\%$ entspricht.

Variable	Maß	Wert	
Alter	Median	21	
	Standardabweichung	3	
	Spanne	$18 - 34$	
	Mittelwert	22	
	Kategorie	**Anzahl**	**Prozent**
Muttersprache	Deutsch	104	81,9
	zweisprachig	6	4,7
	nicht Deutsch	17	13,4
Geschlecht ($n = 128$)	männlich	26	20,3
	weiblich	102	79,7
Fachbereich ($n = 125$)	Erziehungs- und Sozialwissenschaften	25	20,0
	Kulturwissenschaften und ästhetische Kommunikation	2	1,7
	Sprach- und Informationswissenschaften	8	65,6
	Mathematik, Naturwissenschaften, Wirtschaft und Informatik	15	12,0
	Sonstige	1	0,8

dentischen Teilnehmer auf die an der Universität Hildesheim gelehrten Fachbereiche verteilen, ist Tabelle 7.6 zu entnehmen. Hinsichtlich der Teilnehmer aus IT-orientierten Fächern, kann das ursprünglich formulierte Ziel nicht vollständig umgesetzt werden. Insgesamt stammen 82 ($65{,}6\,\%$) der untersuchten Probanden aus IT-orientierten Studiengängen. Davon sind 13 nach eigenen Angaben in einen IT-bezogenen Masterstudiengang eingeschrieben (Angabe nicht obligatorisch). Der Zahl von Studienanfängern an den verbleibenden 69 Teilnehmern aus IT-orientierten Fächern beläuft sich hingegen auf 50 Testpersonen. Da der Anteil der Untersuchungsteilnehmer aus IT-orientierten Fächern, die nicht den Studienanfängern zugerechnet werden können mit insgesamt 32 Teilnehmern relativ hoch ausfällt, wird darüber hinaus ein möglicher Einfluss der Gruppenzugehörigkeit auf die Untersuchungsergebnisse überprüft (vgl. Abschn. 7.4.6).

Tab. 7.7.: Computer- und Sucherfahrung ($n = 127$).

Variable	Median	M	SD	Spanne
Computererfahrung (Jahre)	10	10,3	3,0	$5 - 20$
Computernutzung (Stunden pro Woche)	15	18,7	13,7	$1 - 72$
Suchmaschinenerfahrung (Jahre)	8	8,1	2,1	$3 - 14$
Suchmaschinennutzung (Stunden pro Woche)	3	5,6	5,4	$0{,}5 - 30$
Bekannte Suchmaschinen ($n = 121$)	2	2,5	1,3	$1 - 8$
Verwendete Suchmaschinen ($n = 121$)	1	1,1	0,4	$1 - 4$

Tabelle 7.7 fasst die auf die Computer- und Sucherfahrung der Teilnehmer bezogenen Merkmale der Stichprobe zusammen. Die durchschnittliche Computererfahrung liegt im dritten Experiment bei $10{,}3$ Jahren (SD: $3{,}0$), die durchschnittliche Suchmaschinenerfahrung beträgt $8{,}1$ Jahre (SD: $2{,}1$). Damit liegen beide Werte leicht über denen des zweiten Experiments (Computererfahrung: $9{,}1$ Jahre; Suchmaschinenerfahrung: $6{,}7$ Jahre). In Anbetracht der Tatsache, dass sich die beiden Stichproben bezüglich der Altersverteilung kaum unterscheiden, lässt dies vermuten, dass dieser Jahrgang schon früher mit den Medien Computer und Internet in Berührung gekommen ist. Dies wird zudem deutlich, wenn man das Alter der Probanden einbezieht. Liegt im zweiten Experiment das früheste Alter der Computernutzung noch bei sechs Jahren, so geben aktuell zwei Teilnehmer an bereits im Alter von drei Jahren mit Computern in Kontakt gekommen zu sein.

Die Ergebnisse hinsichtlich der in allen drei Experimenten erfassten Merkmale wöchentliche Computernutzung, Bekanntheit und Nutzung unterschiedlicher Suchdienste sind somit mit denen der ersten beiden Experimente vergleichbar. Jedoch ist erneut ein Trend zu einer erhöhten Computernutzung zu beobachten. So betragen die erhobenen durchschnittlichen Nutzungswerte in der ersten Studie noch $16{,}7$ Stunden pro Woche, während sich dieser Wert für das zweite und dritte Experiment auf $18{,}7$ Stunden erhöht. Wie in den ersten beiden Experimenten ist auch hier die Differenz zwischen Maximum und Minimum relativ groß, was jedoch erneut mit der Studienphase, in der sich die Probanden aktuell befinden, in Zusammenhang stehen könnte. Im Hinblick auf die Bekanntheit und die Nutzung unterschiedlicher Suchdienste zeigt sich in dieser Stichprobe, dass die Befragten im Schnitt $2{,}5$ unterschiedliche Suchmaschinen kennen (SD: $1{,}3$), im Schnitt jedoch nur $1{,}1$ regelmäßig verwenden (SD: $0{,}4$). Die geringen Werte für die Standardabweichungen deuten wie in den ersten beiden Experimenten darauf hin, dass in diesem Fall von einem vergleichbaren Wissensstand ausgegangen werden kann. Anzumerken ist an dieser Stelle, dass fehlerhafte Angaben hinsichtlich der tatsächlich genutzten Suchmaschinen in sechs Fällen dazu führen haben, dass diese Werte bei diesen Probanden fehlen. Die in Tabelle 7.7 dargestellten Werte beziehen sich somit für dieses Merkmal auf ein n von 121 Probanden. Die durchschnittliche Suchmaschinennutzung liegt im dritten Experiment bei $5{,}6$ Stunden pro Woche (SD: $5{,}4$), im Vergleich zu $4{,}3$ Stunden pro Woche im zweiten Experiment (SD: $5{,}2$). Somit ist auch hier bei gleichbleibender Spannweite (R: $29{,}5$) ein leichter Anstieg im Vergleich zum zweiten Experiment zu beobachten.

Tab. 7.8.: Selbsteinschätzung des Suchmaschinenwissens ($n = 127$).

	Suchwissen	
Kategorie	**Anzahl**	**Prozent**
Anfänger	29	22,8
Fortgeschrittene	91	71,7
Experten	7	5,5

In Bezug auf die Selbsteinschätzung ihrer Suchmaschinenkenntnisse verorten sich die Teilnehmer wie im zweiten Experiment zum Großteil in beiden ersten Kategorien (vgl. Tab. 7.8). Während sich die Selbsteinschätzungen der Teilnehmer in der zweiten Nutzerstudie jedoch noch nahezu in gleichem Maße auf die Kategorien Anfänger ($42{,}6\,\%$) und Fortgeschrittene ($55{,}7\,\%$) verteilen, stuft sich die Mehrheit der aktuellen Teilnehmer auf fortgeschrittenem Niveau ein ($71{,}7\,\%$). Dabei ist der höhere Anteil an fortgeschrittenen Teilnehmern nicht auf den erhöhten Anteil von Versuchspersonen aus IT-orientierten Studiengängen ab dem zweiten Semester zurückzuführen, da sich in dieser Gruppe ebenfalls lediglich $75{,}0\,\%$ auf einem fortgeschrittenen Niveau einstufen. Gleiches gilt analog für die zuvor berichteten Erfahrungs- und Nutzungswerte.

Zusammenfassend lässt sich feststellen, dass auch die Stichprobe des dritten Experiments in Bezug auf die soziodemographischen Merkmale als weitgehend homogen bezeichnet werden kann. Kritisch lässt sich erneut das Missverhältnis zwischen weiblichen und männlichen Untersuchungsteilnehmern betrachten, das jedoch wie in Abschnitt 6.4.1 erwähnt, das vorherrschende Verhältnis weiblicher und männlicher Studierender an der Universität Hildesheim widerspiegelt. Darüber hinaus lässt sich auch der relativ hohe Anteil von Teilnehmern aus IT-orientierten Fächern kritisch hinterfragen. Jedoch kann gezeigt werden, dass dieser keinen Einfluss auf die

Sucherfahrung der Teilnehmer auszuüben scheint. Weiterhin zeigt der untersuchungsübergreifende Vergleich der Stichproben, dass die Teilnehmer der drei Experimente in guter Näherung aus einer vergleichbaren Grundgesamtheit stammen. Gleichzeitig ist ein genereller, jahrgangsbedingter Trend zu leicht erhöhten Erfahrungs- und Nutzungswerten zu erkennen.

7.4.2. Auswertungskonzept

Mit dem Ziel auch die zeitliche Abhängigkeit des Erwartungs- und Systemeinflusses auf das Nutzerverhalten zu analysieren, bearbeiten die Probanden im Rahmen der dritten Untersuchung jeweils drei Suchaufgaben unter denselben experimentellen Bedingungen. Pro Testteilnehmer stehen für die Auswertung somit drei Datensätze zur Benutzerleistung und -zufriedenheit zur Verfügung. Auf dieser Datengrundlage werden im Folgenden zwei Auswertungsstrategien verfolgt: Um einen Vergleich mit den Ergebnissen der vorangegangenen Studien zu ermöglichen, werden die Nutzerdaten zunächst über alle drei Aufgaben gemittelt, bevor in einem zweiten Schritt die zeitliche Abfolge der Aufgaben als dritte unabhängige Variable hinzugenommen wird, um auch die dynamische Dimension des Benutzerverhaltens zu berücksichtigen.

Das im Rahmen der dritten Untersuchung angewendete Auswertungskonzept folgt dabei im Grundsatz dem bereits im Kontext von Experiment 2 entwickelten Ansatz, bei dem durch das Ziehen von fünf unabhängigen Zufallsstichproben ausgeglichene Gruppengrößen erreicht werden (vgl. Abschn. 6.4.2). Im Kontext der Mittelwertanalysen stellt dieses Austarieren der Treatmentgruppen das einzige Balancierungskriterium dar, da jeder Teilnehmer im Rahmen der Untersuchung alle drei Aufgaben bearbeitet. Für die dynamische Auswertung der Untersuchungsdaten hingegen, bei der jede Suchaufgabe im Rahmen eines gemischten ANOVA-Designs in die Auswertung einfließt (vgl. Abschn. 4.3.2.4), wird darüber hinaus, im Falle eines vorhandenen Topiceffekts, die Reihenfolge der Aufgabenbearbeitung berücksichtigt. Dieses Vorgehen erscheint gerade vor dem Hintergrund der Auffälligkeiten in Bezug auf das dritte Suchthema (vgl. Abschn. 7.4.1) angemessen, um die interne Validität der Ergebnisse sicherzustellen.

Darüber hinaus findet in Analogie zu Experiment 2 eine qualitative Einteilung der Ausgangsdaten in die Teilmengen SP_A (SP_B u. SP_{UV}) und SP_B statt (vgl. Abschn. 6.4.2 u. 7.4.1), womit zur Einschätzung der Ergebnisse erneut eine konservativere (SP_B) Teilstichprobe einer weniger restriktiven Fallauswahl (SP_A) gegenübergestellt werden kann. Die Mindestgröße zur Berücksichtigung einer unabhängigen Variable wird dabei in Übereinstimmung mit Experiment 2 für SP_B auf 10 und für SP_A auf 20 Teilnehmer pro Untersuchungsgruppe festgelegt. Wie bereits angedeutet erfolgt die Überprüfung der Hypothesen 1 bis 3 (vgl. Abschn. 7.2) dann anhand der über alle drei Aufgaben gemittelten Datensätze $SP_{A,M}$ und $SP_{B,M}$. Die tatsächlich vorliegenden minimalen Stichprobengröße sowie die jeweilige Anzahl verfügbarer Variablen sind in Tabelle 7.9 angegeben.

Zur Untersuchung der vierten Hypothese hingegen, ob Benutzer ihre unrealistischen Erwartungen im Verlauf der Systemnutzung der Systemgüte anpassen, werden, wie eingangs erwähnt, die Auswirkungen von Systemgüte und Erwartungshaltung zusätzlich durch ein Messwiederholungsdesign untersucht, welches den Bearbeitungszeitpunkt als zusätzliche unabhängige Variable berücksichtigt. Da zu diesem Zweck alle Messwerte als Einzelwerte in die Auswertung eingehen und den jeweiligen Messpositionen zugeordnet werden müssen, ist nun im Fall eines Topiceffekts auch eine Balancierung der Aufgabenreihenfolge innerhalb der Versuchsgruppen vorzu-

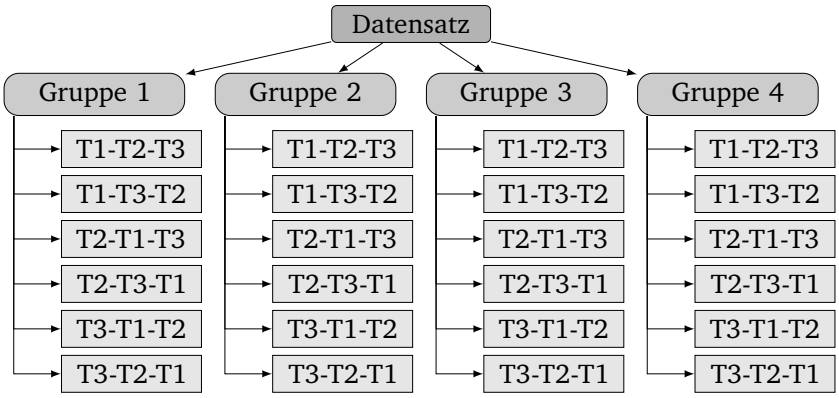


Abb. 7.7.: Vorgehensweise zur Topicbalancierung des Datensatzes nach Versuchsgruppen und Topicreihenfolge. Der Datensatz wird in 24 Treatmentgruppen aufgeteilt, da jede der vier Untersuchungsgruppen zusätzlich bezüglich der sechs möglichen Topicreihenfolgen zerlegt wird.

nehmen. Unter Berücksichtigung der vier Versuchsgruppen sowie der sechs unterschiedlichen Topic-Reihenfolgen ergeben sich somit $4 \times 6 = 24$ Treatments (vgl. Abb. 7.7), auf deren Grundlage die Balancierung vorgenommen wird. Unter Hinzunahme der beiden Datenqualitätsstufen führt dies auf die beiden Teilstichproben $SP_{A,MT}$ und $SP_{B,MT}$. Um auch im Rahmen des dritten Experiments in jedem Fall die größtmögliche Fallzahl pro Untersuchungsgruppe berücksichtigen zu können, werden diese Datensätze um eine weitere Teilmenge (SP_{OT}) ergänzt, welche die abhängigen Variablen ohne Topiceffekt enthält. Um für die einzelnen abhängigen Variablen das Vorliegen eines Topic- oder Reihenfolgeeffekts zu überprüfen, werden einfaktorielle Varianzanalysen mit den Suchthemen als Messwiederholungsfaktor durchgeführt (vgl. Abschn. 7.4.6.1). Hierbei erfolgt die Entscheidung, ob ein Topiceffekt vorhanden ist oder nicht weitestgehend analog zum zweiten Experiment: Ein Topiceffekt wird für eine Variable als eindeutig nicht vorhanden angesehen, wenn jeweils für alle fünf Stichproben von SP_B und SP_A zu keinem der drei Messzeitpunkte ein signifikanter Einfluss eines der Suchthemen beobachtet werden kann.

Da aufgrund der sechs unterschiedlichen Topicreihenfolgen im Fall von $SP_{A,MT}$ und $SP_{B,MT}$ für balancierte Untersuchungsgruppen eine minimale Fallanzahl von zehn Versuchspersonen nicht möglich ist, wird hier die minimale Gruppengröße auf zwölf Teilnehmer festgelegt. Damit ist sichergestellt, dass in jeder der vier Versuchsgruppen in Bezug auf Systemleistung und Erwartungshaltung jede der sechs Topicreihenfolgen mindestens zweimal vorhanden ist. Für die Teilmenge $SP_{B,MT}$ wird dieser Mindeststandard allerdings für keine der unabhängigen Variablen erreicht, weswegen diese Stichprobe für eine Auswertung nicht zur Verfügung steht. Für die Qualitätsstufe $SP_{A,OT}$ wird erneut eine Mindestgröße von 20 und für $SP_{B,OT}$ von 10 Personen pro Gruppe festgesetzt. Wiederum werden aus diesen Teilmengen je fünf Zufallsstichproben gezogen. Tabelle 7.9 fasst die wesentlichen Aspekte aller im dritten Experiment betrachteten Datensätze zusammen.

Zusammenfassend lässt sich sagen, dass auch das für das dritte Experiment angewendete Auswertungskonzept eine optimale Nutzung der Daten ermöglicht, indem eine stark kontrollierte Vorgehensweise mit einer weniger restriktiven Datenauswahl verglichen und in einem einer Kreuzvalidierung ähnlichen Verfahren abgesichert wird.

Tab. 7.9.: Übersicht über verfügbare Datensätze. Neben der Anzahl der unabhängigen Variablen mit ausreichender Fallzahl (Varanz.), sind auch die minimale (n_{min}) sowie die mittlere (n_{mean}) Stichprobengröße angegeben. Der Datensatz $SP_{B,MT}$ ist nicht aufgeführt, da für keine Variable die erforderliche Mindeststichprobengröße erreicht wird. Die angegebenen Zufriedenheitsindikatoren berücksichtigen bereits die acht im Rahmen der Faktorenanalyse entwickelten Zufriedenheitsskalen (vgl. Abschn. 7.4.4). Da für die Auswertung des dynamischen Benutzerverhaltens lediglich die Skalenmittelwerte zur Verfügung stehen, beträgt die Summe der Zufriedenheitsvariablen für $SP_{A,OT}$ und $SP_{A,MT}$ somit 46 im Vergleich zu 54 Variablen für die Mittelwertsauswertung in $SP_{A,M}$.

		Benutzerleistung			Zufriedenheit		
Auswertung	**Datensatz**	**Varanz.**	n_{min}	n_{mean}	**Varanz.**	n_{min}	n_{mean}
Mittelw.	$SP_{B,M}$	129	40	72	54	80	80
Mittelw.	$SP_{A,M}$	115	80	110	54	116	116
Dynamik	$SP_{B,OT}$	37	56	78	5	80	80
Dynamik	$SP_{A,MT}$	77	48	68	41	72	72
Dynamik	$SP_{A,OT}$	37	88	114	5	116	116

7.4.3. Auswertung der Benutzerleistung

In diesem Abschnitt werden die Ergebnisse der Untersuchung in Bezug auf die Suchleistung der Untersuchungsteilnehmer dargestellt. Die in Abschnitt 7.3.2 dargelegten Anpassungen hinsichtlich der Relevanzmessung erlauben im dritten Experiment eine noch differenziertere Betrachtung der Leistungsdaten. Dabei kann zwischen einer fein-abgestuften und einer weniger detaillierten Betrachtungsweise der Relevanz unterschieden werden. Während bei der zuletzt genannten Variante die im ersten und zweiten Experiment verwendete binäre Relevanzskala zugrunde gelegt wird, bezieht die detailliertere Betrachtungsweise alle vier Relevanzstufen mit ein. Im ersten Fall werden bspw. sowohl von den Juroren als *relevant*, als auch als *eher relevant* bewertete Dokumente in die Menge der relevant bewerteten Dokumente aufgenommen. Im zweiten Fall hingegen werden letztere nicht für die Berechnung herangezogen. Wie im vorangegangenen Abschnitt beschrieben, erfolgt die Überprüfung der zweiten und dritten Forschungshypothese (vgl. Abschn. 7.2) zunächst mittels zweifaktorieller Varianzanalysen auf Basis gemittelter Leistungswerte. Eine differenzierte Betrachtung des Mehrwerts der 4-stufigen Relevanzskala findet sich in Abschnitt 7.4.3.2. Um darüber hinaus auch dynamische Effekte der Qualitätsbeurteilung (H4) untersuchen zu können, wird im zweiten Analyseschritt ein Messwiederholungsdesign umgesetzt, bei dem alle Messwerte als Einzelwerte eingehen (vgl. Abschn. 7.4.3.3). Wie bereits in Experiment 2 macht es die Fülle der betrachteten abhängigen Variablen notwendig von einer ausführlichen Diskussion jedes einzelnen Benutzerleistungsmaßes abzusehen. Vielmehr konzentriert sich die folgende Darstellung darauf Beobachtungen, die anhand verschiedener Indikatoren belegt werden können zu bündeln. Dabei erfolgt die Argumentation häufig anhand einiger weniger ausgewählter Leistungsmaße, während weitere, die jeweilige These stützende, Variablen durch Angabe in Klammern dokumentiert werden. Dies ermöglicht zum einen das analytische Vorgehen nachzuvollziehen und andererseits die den Befunden zugrunde liegende Datenbasis zu protokollieren. In diesem Sinne können diese Auflistungen als weiteres Maß der Stabilität bzw. Stärke der jeweiligen Effekte zu interpretiert werden, die zum inhaltlichen Verständnis der Argumentationslinien jedoch nicht notwendigerweise in ihrer Gänze nachvollzogen werden müssen.

7.4.3.1. Varianzanalyse der Mittelwerte

In diesem Abschnitt werden die Ergebnisse der Mittelwertanalyse vorgestellt und diskutiert. Die Tabellen 7.10 bis 7.14 enthalten dabei die Gruppenmittelwerte signifikanter Leistungsmaße dieses ersten Analyseschritts für die Stichproben $SP_{A,M}$ und $SP_{B,M}$. Im Sinne eines Übergangs von allgemeinen zu spezielleren Befunden werden zunächst die in $SP_{A,M}$ nachweisbaren signifikanten Effekte berichtet (Tab. 7.10 u. 7.13), wobei eine Fußnote angibt, ob das entsprechende Ergebnis in $SP_{B,M}$ bestätigt werden kann. Die Tabellen 7.11 und 7.14 hingegen enthalten in der besser kontrollierten Stichprobe $SP_{B,M}$ zusätzlich auftretende Effekte. Die vollständige Dokumentation der Resultate aus Stichprobe $SP_{B,M}$ findet sich hingegen in Anhang E.3. Die Mittelwerte für signifikante Wechselwirkungen zwischen Systemleistung und Erwartungshaltung werden gesammelt in Tabelle 7.12 dargestellt. Wie in den zuvor schon beschriebenen Analysen, werden auch an dieser Stelle ausschließlich Ergebnisse berichtet, die in mindestens vier von fünf Stichproben signifikant sind, wobei für jede Variable die Stichprobe mit dem signifikantesten Ergebnis ausgewählt wird. Falls sich bei zwei signifikanten Haupteffekten die Stichproben mit dem geringsten p-Wert unterscheiden, sind beide Ergebnisse angegeben. Dies ist bspw. bei Z05 in Tabelle 7.11 der Fall. Zur schnelleren Einordnung der Ergebnisse ist die jeweils bessere Suchleistung ebenfalls durch eine Fußnote gekennzeichnet. Anhang E.3 stellt weiterführende Informationen bezüglich der zugrundeliegenden Stichprobengröße, der verwendeten Varianzanalyse (klassisch vs. robust), der Qualität (eindeutig vs. tendenziell) sowie des Signifikanzniveaus der gefundenen Effekte bereit. Im Folgenden werden die Ergebnisse konkreter in Bezug auf die forschungsleitenden Hypothesen diskutiert.

Im Gegensatz zum zweiten Experiment ergeben sich in der dritten Untersuchung sowohl für die Gesamtstichprobe $SP_{A,M}$ als auch für die bereinigte Stichprobe $SP_{B,M}$ signifikante Unterschiede, was vermutlich zum einen die größere Anzahl auswertbarer Leistungsmaße und zum anderen auf die für $SP_{B,M}$ deutlich höheren Stichprobenumfänge im Fall des dritten Experiments zurückzuführen ist (vgl. Abschn. 6.4.2 vs. 7.4.2). Es zeigt sich, dass bei 49 der insgesamt 129 auswertbaren Leistungsvariablen signifikante Mittelwertunterschiede auftreten. Dabei belaufen sich die Stichprobengrößen für $SP_{A,M}$ im Durchschnitt auf 110 und für $SP_{B,M}$ auf 72 Probanden (vgl. Tab. E.21 u. E.22). Insgesamt betrachtet ist die Übereinstimmung zwischen den Ergebnissen beider Datensätze recht hoch. So zeigen z.B. 26 Variablen, für die sich in $SP_{A,M}$ ein signifikanter Effekt nachweisen lässt, diesen ebenfalls in $SP_{B,M}$. Ein ähnliches Maß an Übereinstimmung ergibt sich für die Variablen, die über alle fünf Stichproben hinweg keine signifikanten Unterschiede aufweisen. Hier zeigen 47 Variablen sowohl in $SP_{A,M}$ als auch in $SP_{B,M}$ eindeutig keine signifikanten Effekte, wohingegen für 44 Leistungsmaße dies nur in einer der beiden Teilstichproben der Fall ist (vgl. Tab. E.14).

Der in der zweiten Forschungshypothese postulierte Kompensationseffekt lässt sich im dritten Experiment über beide Stichproben $SP_{A,M}$ und $SP_{B,M}$ nachweisen. So zeigt sich für die Recallmaße V03, V04, V26, V32/BR, V38 und V39 eindeutig kein Einfluss in Bezug auf die Systemqualität (vgl. Tab. E.14). Dies bestätigt und erweitert erneut die Ergebnisse von Al-Maskari et al. (2008b) und Allan et al. (2005) zur Unabhängigkeit der Benutzerleistung von der Systemgüte (vgl. Abschn. 6.4.3 u. 3.2.1. Weiterhin gilt dies für die Variable V27 in der Stichprobe SP_A (vgl. Tab. E.14). Einzig das Recallmaß V33 zeigt in der Stichprobe $SP_{B,M}$ ein anderes Verhalten, nämlich einen in

Tab. 7.10.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei binärer Relevanzskala in $SP_{A,M}$. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen E.15 und E.21 in Anhang E.3 entnommen werden.

		System		Erwartung	
ID	**Beschreibung**	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$
M05[a]	Anz. aufg. irrel. Dok.	**1,50**[b]	**2,69**	1,95	2,24
M07[a]	Anz. falsch irrel. bew. Dok.	**2,38**	**1,65**[b]	1,98	2,05
M08[a]	Anz. falsch rel. bew. Dok.	**0,39**[b]	**0,72**	0,57	0,54
M14[a]	Anz. richtig bew. Dok.	**3,84**	**5,86**[b]	4,66	5,04
M15[a]	Anz. richtig irrel. bew. Dok.	**1,14**	**2,12**[b]	1,49	1,77
B04	Durchschn. Bew. rel. Dok.	**5,29**	**5,73**[b]	5,55	5,47
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	**5,36**	**5,82**[b]	5,68	5,50
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	**42,14**	**37,61**[b]	41,4	38,35
Z02-log[a]	Durchschn. Betrachtungsz. falsch bew. Dok.	**3,48**	**3,33**[b]	3,46	3,35
Z05[a]	Durchschn. Betrachtungsz. irrel. bew. Dok.	**37,42**	**32,99**[b]	36,75	33,66
Z05-log[a]	Durchschn. Betrachtungsz. irrel. bew. Dok.	**3,33**	**3,18**[b]	3,28	3,23
V01[a]	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,15**[b]	**0,28**	0,21	0,22
V02[a]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,85**[b]	**0,72**	0,80	0,77
V05[a]	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,30**	**0,19**[b]	0,23	0,25
V06[a]	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	**0,69**	**0,44**[b]	0,60	0,54
V08[a]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,04**[b]	**0,08**	0,06	0,05
V09[a]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	**0,07**[b]	**0,12**	0,10	0,09
V10[a]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. richtig rel. bew. Dok.}}$	**0,09**[b]	**0,17**	0,13	0,13
V13[a]	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,40**	**0,59**[b]	0,47	0,52
V14[a]	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,11**	**0,21**[b]	**0,14**	**0,18**[b]
V17[a]	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	**0,32**	**0,57**[b]	0,41	0,49
V31/BP[a]	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	**0,94**[b]	**0,90**	0,92	0,92

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.
[b] Dieser Mittelwert entspricht der besseren Benutzerleistung.

der Tendenz erkennbaren Einfluss der Systemleistung. Interessanterweise schneiden hier jedoch die Benutzer des schlechteren Systems besser ab. Sie bewerten also mehr der zurückgegebenen relevanten Dokumente als relevant. Für dieses Verhalten lassen sich zwei mögliche Erklärungen anbringen: Zum einen zeigt sich hier schon der im Folgenden noch genauer erläuterte Anpassungseffekt bezüglich der abnehmenden Übereinstimmungstendenz mit den Jurorenurteilen bei Nutzung des besseren Systems. Zum anderen sind in den Ergebnislisten des schlechteren Systems natürlich weniger relevante Dokumente enthalten. Somit ist es bei gleich hoher Anzahl richtig

relevant bewerteter Dokumente für Benutzer des schlechteren Systems einfacher, eine bessere Leistung in Bezug auf das Leistungsmaß V33 zu erreichen. Abgesehen von diesen beiden Effekten kann somit aber auch für Experiment 3 die zweite Forschungshypothese als bestätigt angesehen werden.

Tab. 7.11.: In $SP_{B,M}$ neu hinzukommende signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei binärer Relevanzskala. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen E.16 und E.22 in Anhang E.3 entnommen werden.

		System		Erwartung	
ID	**Beschreibung**	S_G	S_S	E_H	E_N
M19	Anz. richtig rel. bew. Dok. (letzte Suche)	3,34	3,73	**4,06**[a]	**3,01**
B04	Durchschn. Bew. rel. Dok.	5,33	5,56	**5,68**[a]	**5,21**
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	5,33	5,69	**5,81**[a]	**5,21**
Z01	Durchschn. Betrachtungsz. aller Dok.	**54,64**	**40,27**[a]	44,25	50,65
Z01-log	Durchschn. Betrachtungsz. aller Dok.	**3,64**	**3,44**[a]	3,47	3,61
Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	**37,06** **37,06**	**27,99**[a] **30,26**[a]	33,18 33,84	31,87 33,48
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	**56,91**	**44,74**[a]	47,82	53,84
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	**64,73**	**49,42**[a]	53,39	60,76
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,05**[a] **0,05**[a]	**0,08** **0,07**	**0,08** **0,07**	**0,05**[a] **0,05**[a]
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	0,41	0,41	**0,37**[a]	**0,45**
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	0,59	0,58	**0,63**[a]	**0,54**
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,11**	**0,23**[a]	0,16	0,19
V16	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. falsch irrel. bew. Dok.}}$	**0,51**	**1,08**[a]	0,87	0,72
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	**0,66**	**0,77**[a]	0,74	0,69
V33	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	**0,08**	**0,11**[a]	0,10	0,09

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Bevor im Folgenden auf die Unterschiede zwischen den beiden Datensätzen eingegangen wird, werden zunächst signifikante Ergebnisse besprochen, die über beide Stichproben stabil bleiben und somit die zuverlässigsten Aussagen erlauben. Von den 26 Leistungsmaßen, für die dies der Fall ist, entsprechen die meisten Ergebnisse den Erwartungen. So führt die Nutzung des besseren Systems in den meisten Fällen zu einer höheren Benutzerleistung. Für die untersuchten Dokumentenmengen bedeutet dies bspw., dass, wie im zweiten Experiment, im Fall des besseren Systems weniger irrelevante Dokumente aufgerufen (M05, 4-st.: M20 u. M22) oder falsch als relevant bewertet werden (M08). Außerdem werden für viele der untersuchten Precision- (V02 u. V31/BP) und Imprecisionmaße (V01, V08 u. V09, 4-st.: V36) bei höherer Systemleistung bessere Leistungswerte erzielt.

Auch der durch die Systemleistung hervorgerufene Anpassungseffekt wird im dritten Experiment erneut bestätigt. Besonders ausgeprägt, d. h. in beiden Stichproben ($SP_{A,M}$ u. $SP_{B,M}$)

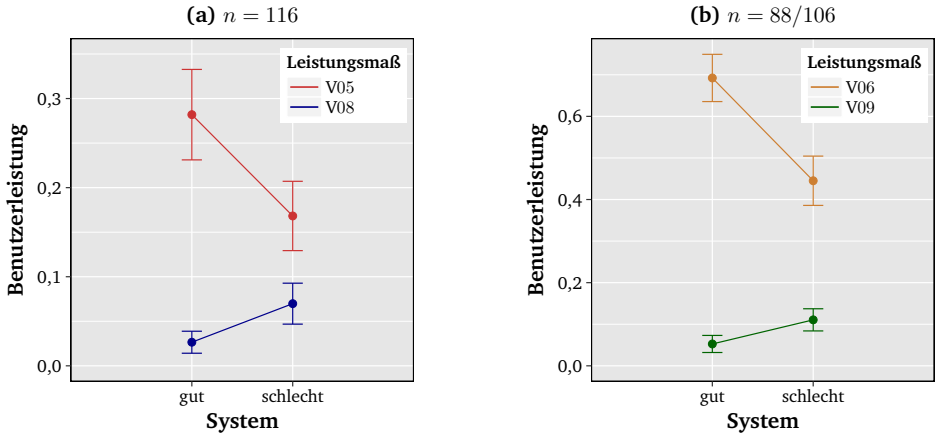


Abb. 7.8.: Systembedingte Anpassung der Relevanzwahrnehmung im dritten Experiment. Dargestellt sind die Anteile der falsch irrelevant und falsch relevant bewerteten Dokumente in Bezug auf die aufgerufenen Dokumente (a) sowie in Bezug auf die relevant bzw. irrelevant bewerteten Dokumente (b). Im Vergleich zum schlechten System ergibt sich für das gute System jeweils eine Zunahme der falsch irrelevant bewerteten Dokumente (V05 u. V06) bzw. eine Abnahme der falsch relevant bewerteten Dokumente (V08 u. V09), wobei dieser Effekt geringer ausfällt. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

signifikant, ist dieser Effekt bei den Variablen M07, M14, M15, V05, V06, V13, V14 und V17 sowie den 4-stufigen Leistungsmaßen M42, M44 und V61. Für alle Variablen zeigen Benutzer des schlechteren Systems eine bessere Leistung. Um die Darstellung nicht zu überfrachten und da die wesentliche Argumentation für alle Maße ähnlich und bereits in Abschnitt 6.4.3 vorgestellt ist, wird der Anpassungseffekt im Folgenden nur anhand einiger ausgewählter, charakteristischer Variablen näher erläutert sowie grafisch in Abbildung 7.8 dargestellt. So zeigt sich bspw., dass insgesamt die Übereinstimmungstendenz zwischen Juroren- und Benutzerurteil mit steigender Systemleistung sinkt (V13). Auf den ersten Blick überrascht dieses Ergebnis, da auf der anderen Seite die Benutzerprecision (V31/BP), also der Anteil der in Übereinstimmung mit den Juroren als relevant bewerteten Dokumente an allen als relevant bewerteten Dokumenten, mit steigender Systemleistung zunimmt. Dieser scheinbare Widerspruch lässt sich jedoch leicht unter Einbezug der weiteren signifikanten Leistungsmaße auflösen. Zwar nimmt die Tendenz mit einem positiven Jurorenurteil übereinzustimmen bei Benutzern des besseren Systems tatsächlich ab (V06), gleichzeitig jedoch neigen diese Probanden weniger dazu, irrelevante Dokumente als relevant zu akzeptieren (V09) (vgl. Abb. 7.8). Insgesamt überwiegt dann der letztere Effekt, was in der Summe zu der beobachteten verbesserten Benutzerprecision führt. Unabhängig davon, können diese Beobachtungen in Bezug auf V09 und V06 sowie die übrigen oben erwähnten Leistungsmaße wieder als Anpassungsreaktion der Relevanzkriterien an die vorgefundene Systemqualität verstanden werden, infolge derer Benutzer des besseren Systems eine restriktivere Bewertungsstrategie verfolgen als Benutzer, die der geringeren Systemqualität ausgesetzt sind. In diesem Sinne lässt sich die erhöhte Benutzerprecision (V31/BP) nun ebenfalls als Indiz für die Anpassung der Relevanzkriterien interpretieren, da sie anzeigt, dass Benutzer des besseren Systems weniger irrelevante Dokumente als relevant akzeptieren. Somit kann die dritte Forschungshypothese als bestätigt angesehen und davon ausgegangen werden, dass Benutzer ihre Relevanzdefinition der

Systemgüte anpassen.

Interessant sind in diesem Zusammenhang auch die nachgewiesenen Effekte der Systemleistung auf die Effizienz bzw. Schnelligkeit der Benutzer. In beiden Stichproben finden sich signifikante Unterschiede in Bezug auf die Betrachtungsdauern unterschiedlicher Dokumentenmengen. Sowohl die benötigte Zeit, um ein Dokument als irrelevant zu bewerten (Z05/Z05-log), als auch die Zeit, die beansprucht wird, um ein Dokument im Widerspruch zu den Juroren zu bewerten (Z02-log) steigt mit der Systemleistung an. Zwar erscheinen die absoluten Differenzen mit gut 4s nicht zu groß auszufallen, jedoch ist zu beachten, dass dies im Vergleich mit den Gesamtbetrachtungszeiten (im Mittel 30s) einem relativen Unterschied von etwa 15 % entspricht. Folglich scheinen Benutzer des besseren Systems mehr Zeit zu benötigen, um die zuvor beobachteten strengeren Relevanzkriterien anzuwenden. Mit anderen Worten deutet dies darauf hin, dass die Relevanzentscheidungen wohlüberlegt und nicht leichtfertig getroffen werden.

Die bisher im Kontext der Systemleistung beschriebenen Ergebnisse werden noch durch eine Reihe weiterer Leistungsmaße gestützt, wenngleich in diesen Fällen die Stabilität bezüglich $SP_{A,M}$ und $SP_{B,M}$ nicht gegeben ist. Zu nennen sind in diesem Zusammenhang zum einen die 4-stufigen Maße M26 und M31 ($SP_{A,M}$) sowie V16 und V29 ($SP_{B,M}$), die den Anpassungseffekt bestätigen. Zum anderen werden noch weitere Zeitmaße signifikant (Z02 ($SP_{A,M}$) sowie Z01/Z01-log, Z09, Z11, 4-st: Z15 ($SP_{B,M}$)). Bemerkenswert ist in diesem Zusammenhang, dass einige dieser Zeitmaße, wie bspw. Z09, Z11 und Z15, den beobachteten Effizienzunterschied auf einer noch grundsätzlicheren Ebene erklären. So zeigt sich, dass Benutzer des besseren Systems generell, also auch für in Übereinstimmung mit den Juroren als relevant bewertete Dokumente, mehr Zeit benötigen, um ihre Relevanzentscheidungen zu treffen. Zusammenfassend können also auch in der dritten Nutzerstudie die Forschungshypothesen im Hinblick auf Kompensationseffekte für recallorientierte Benutzerleistungsmaße sowie für das Auftreten eines systemleistungsbedingten Anpassungseffekts der Relevanzwahrnehmung bestätigt werden.

Im Gegensatz zu Haupteffekten der Systemleistung sind Haupteffekte der Erwartungshaltung nicht durchgehend stabil, sondern treten zum überwiegenden Teil nur in der bereinigten Stichprobe $SP_{B,M}$ auf. Dabei lassen sich drei Gruppen von Effekten unterscheiden: Effekte, die nur in der Stichprobe $SP_{B,M}$ auftauchen (M19, V11 u. V12, 4-st.: B18, u. V54), Effekte, die in einer der beiden Stichproben zu einem signifikanten Systemeffekt hinzukommen (V14 ($SP_{A,M}$) sowie V08 ($SP_{B,M}$)) und Effekte, die einen in $SP_{A,M}$ signifikanten Systemeffekt durch einen signifikanten Erwartungseffekt in $SP_{B,M}$ ablösen (B04 u. B06, 4-st.: B16). Die Tatsache, dass Effekte der Erwartungshaltung fast nur unter den sehr kontrollierten Bedingungen von Stichprobe $SP_{B,M}$ auftreten, in der viele mögliche Störeinflüsse unterdrückt werden, zeigt wiederum die Schwierigkeit, den Einfluss von Erwartungen im Rahmen eines realistischen Suchszenarios messbar zu machen.

In der Gruppe der ausschließlich in $SP_{B,M}$ auftretenden Effekte zeigt sich zunächst ein weiterer, in diesem Fall jedoch durch die Erwartungshaltung hervorgerufener, Anpassungseffekt der Relevanzbewertung. Nutzer mit einer positiven Voreinstellung bewerten einen signifikant höheren Anteil der von ihnen aufgerufenen Dokumente als relevant (V12). Im Umkehrschluss führt dies zu einem geringeren Anteil von Dokumenten, die als irrelevant verworfen werden (V11, 4-st.: V54). Diese erhöhte Zustimmungstendenz spiegelt sich zudem in den spezielleren Leistungsmaßen M19 und B18 (4-st.) wider. Derselbe Anpassungseffekt lässt sich außerdem noch in der

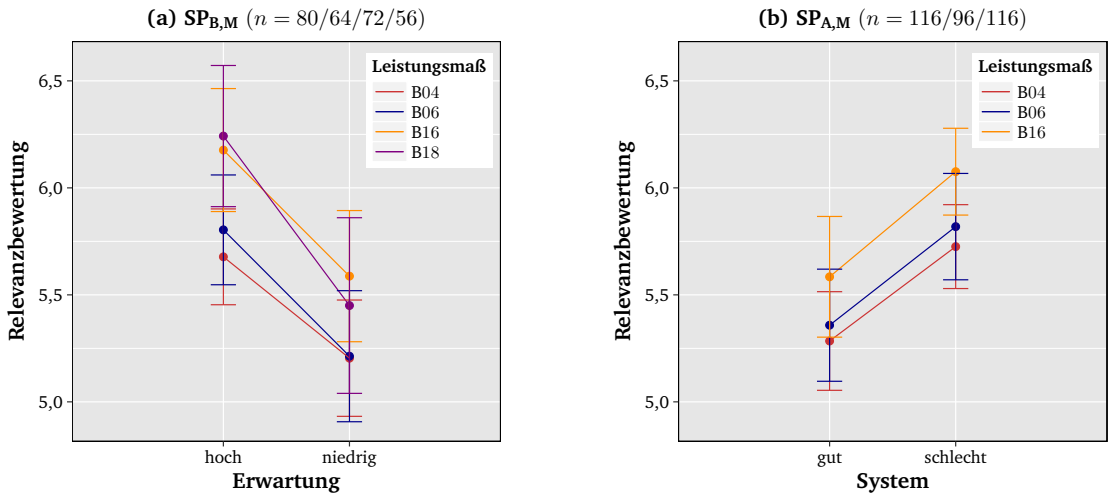


Abb. 7.9.: Übergang von einer erwartungsbedingten Anpassung der Relevanzwahrnehmung in SP$_{B,M}$ zu einer systembedingten Anpassung der Relevanzwahrnehmung in SP$_{A,M}$. Bild (a): Eine höhere Erwartungshaltung führt bei SP$_{B,M}$ zu einer positiveren Bewertung relevanter Dokumente sowohl in Bezug auf die binäre (B04) als auch in Bezug auf die 4-stufige Relevanzskala (B16). Schränkt man die Betrachtung auf die letzte durchgeführte Suche ein, scheint dies tendenziell eine leichte Zunahme des Effekts zu bewirken (B04 vs. B06 u. B16 vs. B18). Bild (b): Der signifikante Einfluss der Erwartungshaltung ist in SP$_{A,M}$ nicht sichtbar. Für drei Leistungsmaße (B04, B06 u. B16) wird nun hingegen die Systemleistung signifikant und zeigt die bereits beschriebene strengere Relevanzbewertung im Falle einer erhöhten Systemleistung. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte. Mittel- und p-Werte für die 4-stufigen Relevanzskalen (B16 u. B18) sind in Abschnitt 7.4.3.2 angegeben.

Gruppe der Leistungsmaße, bei denen beide Haupteffekte signifikant werden, nachweisen. In diesen Fällen (V08 u. V14) wird die Relevanz der Dokumente von Probanden mit der positiveren Voreinstellung zum System ebenfalls höher bewertet.

Besonders interessant erscheinen die Fälle, in denen sich ein in SP$_{B,M}$ signifikanter Erwartungseffekt für SP$_{A,M}$ in einen Systemeffekt transformiert (vgl. Abb. 7.9). Dies ist für die drei Variablen B04, B06 und B16 (4-st.) der Fall, die alle unterschiedliche Aspekte der durchschnittlichen Bewertung relevanter Dokumente erfassen. Während hier in SP$_{B,M}$ erneut die zuvor beschriebene mildere Relevanzbewertung bei hoher Erwartung zu beobachten ist, wandelt sich diese in der weniger kontrollierten Stichprobe SP$_{A,M}$ zu der strengeren Relevanzbewertung bei hoher Systemleistung um. Dies ist umso interessanter, als dass diese drei Leistungsmaße im Gegensatz zu bspw. dem Anteil falsch bewerteter Dokumente, explizit die Relevanzwahrnehmung relevanter Dokumente abfragen. Die Tatsache, dass signifikante Erwartungseinflüsse ausschließlich in SP$_{B,M}$ auftreten, kann darüber hinaus, wie bereits angedeutet, als weiterer Hinweis darauf gewertet werden, dass sich diese Effekte in Bezug auf die Benutzerleistung nur in einem sehr gut kontrollierten experimentellen Umfeld nachweisen lassen. Dabei ist außerdem bemerkenswert, dass sich die Richtung des Anpassungseffektes von Systemleistung zu Erwartungshaltung umkehrt. In Bezug auf die Erwartungshaltung lässt sich aber in allen Fällen zusammenfassend festhalten, dass eine positive Einstellung zum System zu einer weniger strengen Relevanzbewertung führt.

In insgesamt drei Fällen (M02, M17 u. Z05) treten signifikante Wechselwirkungen auf, deren Mittelwerte Tabelle 7.12 entnommen werden können. Besonders deutlich und stabil ist dieser

Tab. 7.12.: Signifikante Interaktionseffekte der Varianzanalysen zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei binärer und 4-stufiger Relevanzskala in $SP_{A,M}$ und $SP_{B,M}$. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu p-Werten können Tabelle E.21 und E.22 in Anhang E.3 entnommen werden.

				Interaktion			
		ID	Beschreibung	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
$SP_{A,M}$	binär	M02	Anz. aufg. Dok. (erste 10 Dok.)	**3,47**	**2,84**	**2,35**	**3,30**
	4-st.	V40[a]	Anz. falsch eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	**0,91**	**0,77**	**0,61**	**0,80**
		V57[a]	Anz. richtig eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	**0,06**	**0,29**	**0,36**	**0,20**
$SP_{B,M}$	binär	M02	Anz. aufg. Dok. (erste 10 Dok.)	**3,83**	**2,47**	**2,69**	**3,36**
		M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	**2,11**	**1,31**	**1,33**	**1,53**
		Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	**42,11**	**34,26**	**23,12**	**30,70**
	4-st.	M27	Anz. falsch eher irrel. bew. eher rel. Dok.	**0,86**	**0,41**	**0,28**	**0,56**
		V40[a]	Anz. falsch eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	**0,93**	**0,79**	**0,59**	**0,84**

[a] Stichprobengröße < 40.

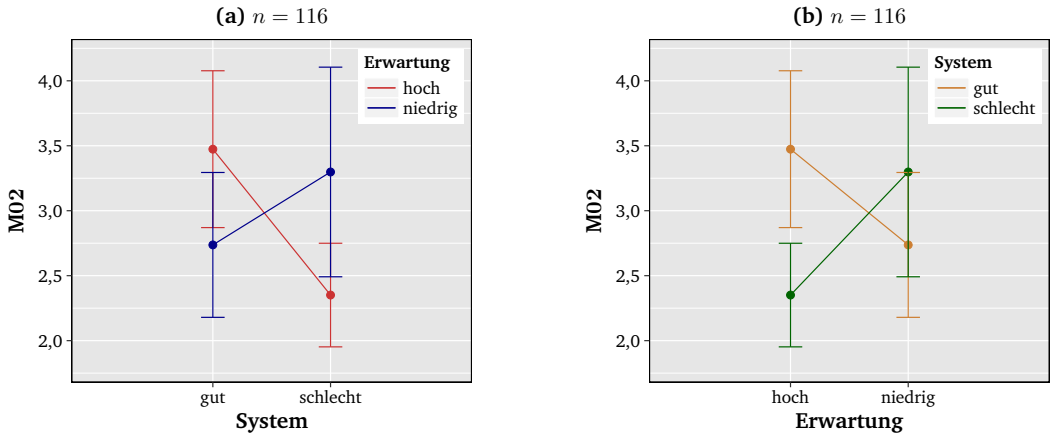


Abb. 7.10.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für die Anzahl der aufgerufenen Dokumente der ersten 10 angezeigten Dokumente (M02) in SP_A. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bei hoher Erwartungshaltung steigt die Anzahl aufgerufener Dokumente bei der guten im Vergleich zur schlechten Systemleistung an. Bei niedriger Erwartungshaltung scheint sich dieser Effekt tendenziell umzukehren. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

Effekt für die Anzahl der aufgerufenen Dokumente der ersten zehn angezeigten Dokumente (M02), der auch unter Einbeziehung problematischer Fallgruppen ($SP_{A,M}$) erhalten bleibt. Das gleiche Verhalten zeigt sich darüber hinaus für die richtig als relevant bewerteten Dokumente derselben Dokumentenmenge (M17), wobei hier der Effekt ausschließlich für $SP_{B,M}$ zutage tritt. Mithilfe der in Abbildung 7.10 beispielhaft für M02 dargestellten Mittelwertunterschiede lassen sich beide Interaktionseffekte folgendermaßen interpretieren: Betrachtet man den Einfluss der Er-

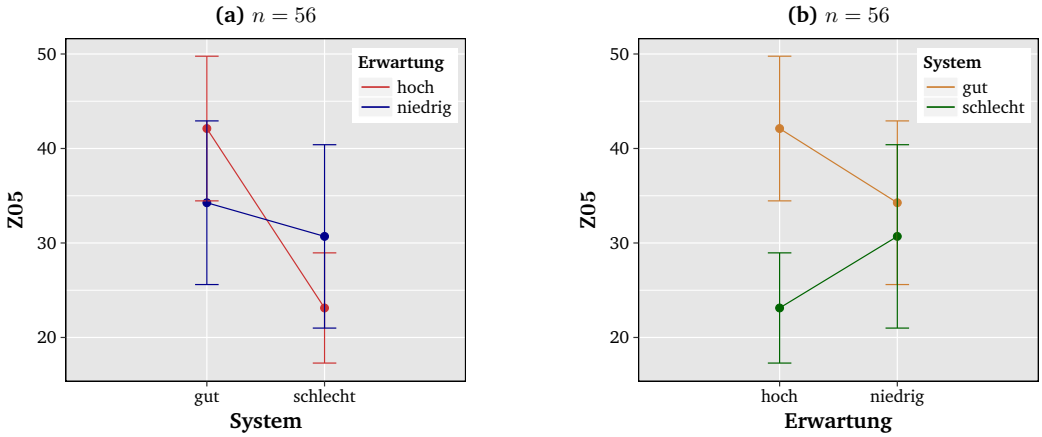


Abb. 7.11.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für die durchschnittliche Betrachtungszeit irrelevant bewerteter Dokumente (Z05) in SP_B. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bei niedriger Erwartungshaltung lässt sich im Rahmen der Fehlerbalken kein Unterschied in der Bewertungszeit in Bezug auf die Systemleistung feststellen. Bei hoher Erwartungshaltung hingegen fällt die Bewertungszeit bei guter Systemleistung höher aus als bei niedriger Systemleistung. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

wartungshaltung in Abhängigkeit der Systemleistung, rufen die Nutzer des schlechteren Systems bei hoher Erwartungshaltung weniger Dokumente auf (M02) bzw. bewerten weniger Dokumente in Übereinstimmung mit den Juroren als relevant (M17), als bei niedriger Erwartungshaltung. Im Zusammenhang mit dem besseren System kehrt sich dieses Verhalten tendenziell um. Betrachtet man umgekehrt den Einfluss der Systemgüte in Abhängigkeit der Erwartungshaltung (Bild (b)), ist derselbe Zusammenhang erkennbar: Im Kontext der hohen Erwartungshaltung zeigt sich ein deutlicher Unterschied zwischen den beiden Systemgüten, mit geringerer Anzahl aufgerufener Dokumente, wenn die präsentierte Systemgüte hinter den Erwartungen zurückbleibt. Umgekehrt ist bei niedriger Erwartungshaltung tendenziell erneut das gegenteilige Verhalten zu beobachten. Die Tatsache, dass sich diese Effekte gerade hinsichtlich der ersten zehn angezeigten Dokumente manifestieren, kann als weiteres Indiz für die im zweiten Experiment vermutete dynamische Abhängigkeit des Erwartungseinflusses gewertet werden.

Die signifikante Interaktion zwischen Systemgüte und Erwartungshaltung in Bezug auf Z05 liefert darüber hinaus einen vertieften Einblick in die weiter oben beschriebene Schwierigkeit, bei Nutzung des besseren Systems Dokumente als irrelevant zu verwerfen. Wie in Abbildung 7.11 zu sehen, kommt der Unterschied zwischen den Systemleistungen im Wesentlichen dadurch zustande, dass Benutzer mit einer hohen Erwartung im Fall des besseren Systems deutlich länger für ihre Relevanzentscheidung benötigen, als im Fall des schlechteren Systems. Für Probanden mit der niedrigen Erwartungshaltung fällt dieser Unterschied hingegen geringer aus. Somit scheint der beschriebene Unterschied beim Verwerfen als irrelevant wahrgenommener Dokumente vor allem im Kontext einer hohen Erwartung aufzutreten.

7.4.3.2. Vierstufige Relevanzskala

Als abschließende Betrachtung der gemittelten Leistungsmaße wird im Folgenden spezieller auf ausgewählte Ergebnisse eingegangen, die sich im Kontext der im dritten Experiment neu eingeführten 4-stufigen Relevanzskala ergeben (vgl. Abschn. 7.3.4). Im Prinzip werden dabei als Leistungsmaße die gleichen Dokumentenmengen und Verhältnisse betrachtet wie im binären Fall, allerdings ergeben sich durch die Vierstufigkeit einige zusätzliche Untermengen. So kann bspw. die Menge der falsch als eher relevant bewerteten Dokumente nun in drei Untermengen zerlegt werden, je nachdem, ob es sich um eigentlich irrelevante (M32), eher irrelevante (M31) oder relevante (M33) Dokumente handelt.

Tab. 7.13.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei 4-stufiger Relevanzskala in $SP_{A,M}$. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen E.18 und E.21 in Anhang E.3 entnommen werden.

		System		Erwartung	
ID	**Beschreibung**	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$
M20[a]	Anz. aufg. eher irrel. Dok.	**0,93**[b]	**1,72**	1,24	1,41
M22[a]	Anz. aufg. irrel. Dok.	**0,52**[b]	**1,11**	0,78	0,85
M26	Anz. falsch eher irrel. bew. Dok.	**1,50**	**1,21**[b]	1,32	1,39
M31[a]	Anz. falsch eher rel. bew. eher irrel. Dok.	**0,18**[b]	**0,36**	0,33	0,21
M42[a]	Anz. richtig eher irrel. bew. Dok.	**0,23**	**0,46**[b]	0,34	0,35
M44[a]	Anz. richtig irrel. bew. Dok.	**0,36**	**0,77**[b]	0,56	0,57
B16	Durchschn. Bew. rel. Dok.	**5,59**	**6,08**[b]	5,93	5,74
V36[a]	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,05**[b]	**0,10**	0,07	0,08
V37	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,50**[b]	**0,45**	0,49	0,45
V45[c]	$\frac{\text{Anz. falsch eher rel. bew. eher irrel. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	**0,06**[b]	**0,15**	0,13	0,08
V40[a,c]	$\frac{\text{Anz. falsch eher irrel. bew. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	**0,83**	**0,69**[b]	0,76	0,77
V57[a]	$\frac{\text{Anz. richtig eher irrel. bew. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	**0,16** **0,18**	**0,29**[b] **0,29**[b]	0,22 0,21	0,23 0,26
V58[c]	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	0,36	0,32	**0,31**	**0,38**[b]
V61[a]	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,04**	**0,07**[b]	0,05	0,06

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.
[b] Dieser Mittelwert entspricht der besseren Benutzerleistung.
[c] Stichprobengröße < 80.

In diesem Zusammenhang ist auch zu beachten, dass diese zusätzliche Zerlegung zu einer weiteren Verringerung der Fallzahlen pro Untersuchungsgruppe führen kann. Um trotzdem an dem gewählten Auswertungskonzept bezüglich $SP_{A,M}$ und $SP_{B,M}$ festhalten zu können, ist es deshalb bei einigen Maßen notwendig die bisher verwendete Mindestgruppengröße von 10 bzw. 20 Testpersonen aufzuheben. Dies ist für die Variablen V40, V41, V42, V43, V49, V57, V64, V78, V79 und V80 in Bezug auf beide Stichproben $SP_{A,M}$ und $SP_{B,M}$ sowie für die Leistungsmaße V44, V45, V46, V47 und V58 in Bezug auf $SP_{A,M}$ der Fall. Der minimale Stichprobenumfang sinkt damit auf

insgesamt 28 Testpersonen für $SP_{B,M}$ und 36 Probanden für $SP_{A,M}$. In dieser Hinsicht ist die Aussagekraft der Ergebnisse also jeweils vor dem Hintergrund der tatsächlich vorhandenen Fallzahlen zu sehen. Die Resultate dieser Analyse sind in den Tabellen 7.13 und 7.14 zusammengefasst. Es können signifikante Haupteffekte der Systemleistung, der Erwartungshaltung sowie Interaktionen nachgewiesen werden. Im Ergebnis erlauben die Resultate einen vertieften Einblick in die im Kontext der binären Relevanzskala beschriebenen Effekte.

Tab. 7.14.: In $SP_{B,M}$ neu hinzukommende signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei 4-stufiger Relevanzskala. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen E.17 und E.22 in Anhang E.3 entnommen werden.

		System		Erwartung	
ID	**Beschreibung**	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$
B16	Durchschn. Bew. rel. Dok.	5,74	6,03	**6,18**[a]	**5,59**
B18	Durchschn. Bew. rel. Dok. (letzte Suche)	5,69	6,01	**6,24**[a]	**5,45**
Z15	Durchschn. Betrachtungsz. eher rel. Dok.	**50,99**	**35,64**[a]	43,02	43,61
V41[b]	$\frac{\text{Anz. falsch eher irrel. bew. eher rel. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	**0,44**	**0,24**[a]	0,42	0,26
V54	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	0,20	0,23	**0,18**[a]	**0,25**

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.
[b] Stichprobengröße < 40.

Zunächst zeigt sich, dass auch für die feinere Relevanzskala Systemunterschiede in Bezug auf die Recallmaße vollständig kompensiert werden können. Dies gilt sowohl für die Recallmaße der relevanten (V38, V39, V79 u. V80,) wie auch für die Recallmaße der eher relevanten Dokumente (V34, V35, V59 u. V60). Dabei ist es unerheblich ob die aufgerufenen oder lediglich die richtig bewerteten Dokumente betrachtet werden oder ob als Bezugsmenge die entsprechenden Dokumente im Korpus oder ausschließlich die tatsächlich zurückgegebenen Dokumente zugrunde gelegt werden.

Des Weiteren erlaubt die 4-stufige Skala eine differenziertere Analyse der zuvor beschriebenen Anpassungseffekte. Dabei zeigt sich als erstes interessantes Ergebnis, dass beide Anpassungseffekte eher im Bereich der eher relevanten und eher irrelevanten Dokumente zu verorten sind. Die relevanten und irrelevanten Dokumente dagegen sind weitestgehend nicht betroffen. Deutlich wird dies anhand der Tatsache, dass die Anteile richtig bzw. falsch bewerteter Dokumente sowohl für die relevanten als auch für die irrelevanten Dokumente in mindestens einer der beiden Stichproben $SP_{A,M}$ oder $SP_{B,M}$ sowohl für die Erwartung als auch für die Systemleistung eindeutig nicht signifikant werden (V78, V64, V49, V52). Für die beiden Leistungsmaße V78 und V52 gilt dies sogar für beide Stichproben. Für V49 hingegen lässt sich in Bezug auf die Stichprobe $SP_{B,M}$ keine eindeutige Aussage treffen, während dies bei V64 für die Stichprobe $SP_{A,M}$ der Fall ist. Im Gegensatz dazu können jedoch eindeutig signifikante Effekte für die mittleren Bewertungskategorien nachgewiesen werden. So ergibt sich für den Anteil der richtig als eher relevant bewerteten Dokumente (V58) ein signifikanter Effekt der Erwartungshaltung für die Stichprobe $SP_{A,M}$. Darüber hinaus kann für die richtig als eher irrelevant bewerteten Dokumente

(V57) ein in Bezug auf $SP_{A,M}$ und $SP_{B,M}$ stabiler systembedingter Anpassungseffekt nachgewiesen werden. Dies gilt gleichermaßen für den Anteil der falsch als eher irrelevant bewerteten Dokumente (V40), wobei für die beiden letztgenannten Leistungsmaße zusätzlich die im Folgenden beschriebenen signifikanten Interaktionen zwischen Systemgüte und Erwartungshaltung für die Stichprobe $SP_{A,M}$ zu beachten sind. Einzig für den Anteil der fälschlicherweise als eher relevant bewerteten Dokumente kann kein eindeutig oder in der Tendenz vorliegender Anpassungseffekt beobachtet werden (V44). Zusammenfassend scheinen diese Ergebnisse jedoch darauf hinzudeuten, dass beide Anpassungseffekte eher im Bereich der mittleren Bewertungskategorien zu verorten sind.

Deutlich wird der Mehrwert der 4-stufigen Skala zudem an den signifikanten Interaktionseffekten von M27, V40 und V57. Dabei lässt sich diese Interaktion für V57 in $SP_{A,M}$ und für M27 in der besser kontrollierten Stichprobe $SP_{B,M}$ nachweisen. Das Leistungsmaß V40 hingegen zeigt diesen Effekt stabil über beide Stichproben. Die entsprechenden Gruppenmittelwerte können Tabelle 7.12 und Abbildung 7.12 entnommen werden. Da sich alle drei Leistungsmaße auf die als eher irrelevant bewerteten Dokumente beziehen, zeigt sich auch hier, dass die schon im binären Kontext beobachtete Abnahme der Zustimmungstendenz zu den Jurorenurteilen bei steigender Systemleistung tatsächlich bezüglich der mittleren Bewertungskategorien entsteht. Dabei zeigen V40 und V57, als Anteil der falsch bzw. richtig als eher irrelevant bewerteten Dokumente, wie erwartet ein gegenläufiges Verhalten. Wie in Abbildung 7.12 zu sehen, ist die Abnahme der Zustimmungstendenz nicht allein an die Systemleistung gekoppelt. Vielmehr wird der Einfluss der Systemleistung durch die Erwartung der Probanden moderiert. Während es bei Benutzern mit der niedrigeren Erwartungshaltung zu keiner nennenswerten Adaption der Relevanzkriterien in Abhängigkeit von der Systemleistung kommt, ändert sich dies bei Benutzern mit der höheren Erwartung. Hier führt eine hohe Systemleistung zu einer deutlich strengeren und die niedrige Systemleistung zu einer deutlich weniger restriktiven Relevanzbeurteilung. Die Anpassung der Relevanzkriterien in Abhängigkeit der Erwartungshaltung scheint damit also an eine hohe Erwartungshaltung der Testteilnehmer gekoppelt zu sein. Interessant ist hierbei zudem, dass die resultierenden Werte höher bzw. niedriger als im Fall der niedrigen Erwartungshaltung liegen. Die Anpassung findet also gegenläufig für beide Systemqualitäten statt. Eine offensichtliche Frage, die sich an dieser Stelle ergibt, ist ob relevante, eher relevante oder irrelevante Dokumente von Benutzern des besseren Systems fälschlicherweise als eher irrelevant bewertet werden, ob also tatsächlich von einer strengeren Relevanzbewertung gesprochen werden kann. Zusammen mit den eindeutig nicht signifikanten Ergebnissen in Bezug auf M29 und M28 deutet die signifikante Wechselwirkung in Bezug auf M27 darauf hin, dass dies tatsächlich der Fall sein könnte. Die relevanten und irrelevanten Dokumente, die fälschlicherweise als eher irrelevant bewertet werden, sind eindeutig bzw. in der Tendenz nicht signifikant (M29 u. M28). Hingegen zeigt die Menge der eher relevanten Dokumente, die fälschlicherweise als eher irrelevant bewertet werden (M27) den zu V40 und V57 analogen Anpassungseffekt (vgl. Abb. 7.12). Allerdings ist zu beachten, dass obwohl die Ergebnisse in der Gesamtschau eine stringente Interpretation erlauben, die Interaktion für M27 eindeutig nur in $SP_{B,M}$ die Wechselwirkungen für V40 und V57 hingegen eindeutig nur in $SP_{A,M}$ nachgewiesen werden können. Vor dem Hintergrund dieser Erwartungsabhängigkeit des Systemeinflusses lässt sich nun auch die in Experiment 2 beobachtete Wechselwirkung in Bezug

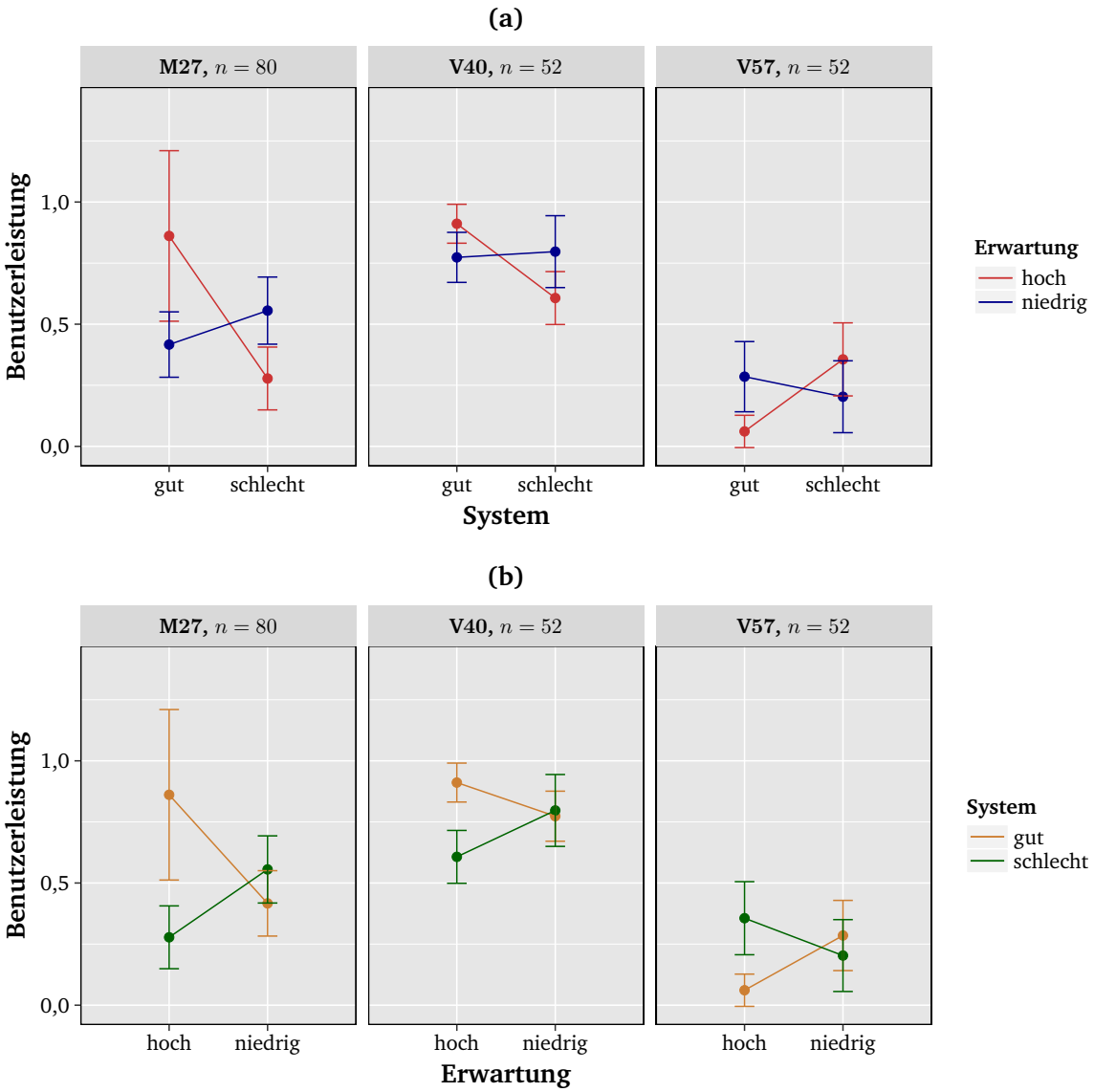


Abb. 7.12.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für die Anteile der richtig (V57) und falsch (V40) als eher irrelevant bewerteten Dokumente sowie für die Anzahl der eher relevanten Dokumente, die fälschlicherweise als eher irrelevant bewertet wurden (M27). Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bei niedriger Erwartungshaltung lässt sich im Rahmen der Fehlerbalken kein Unterschied zwischen den beiden Systemgüten feststellen. Bei hoher Erwartungshaltung hingegen fällt der Anteil der fälschlicherweise als eher irrelevant bewerteten Dokumente für das bessere System höher aus als für die schlechtere Systemleistung (M27 u. V40). Für den Anteil richtig als eher irrelevant bewerteter Dokumente ergibt sich entsprechend der umgekehrte Effekt (V57). Im Zusammenhang mit den nicht signifikanten Ergebnissen in Bezug auf die Anzahl relevanter (M29) und irrelevanter (M28) Dokumente die fälschlicherweise als eher irrelevant bewertet wurden, deutet dies erneut auf die Anwendung strengerer Relevanzkriterien im Zusammenhang mit einer besseren Systemleistung hin, die allerdings nur im Kontext einer hohen Erwartungshaltung zum Tragen kommt und sich darüber hinaus im Wesentlichen auf der Ebene der mittleren Bewertungskategorien abspielt. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

auf das Leistungsmaß M02 neu interpretieren (vgl. Abb. 6.14). Die Anpassung der Relevanzwahrnehmung im Kontext der hohen Erwartungshaltung führt dazu, dass die wahrgenommene Qualität für beide Suchsysteme gleich hoch ausfällt und aus diesem Grund eine ähnliche Anzahl von Dokumenten aufgerufen wird. Im Fall der niedrigen Erwartungshaltung hingegen realisieren die Probanden den Systemunterschied und rufen im Fall der höheren Systemqualität eine größere Anzahl von Dokumenten auf.

Das Leistungsmaß V58, das den Anteil richtig als eher relevant bewerteter Dokumente beschreibt, zeigt in der Tendenz eine signifikante Erwartungsabhängigkeit für die Stichprobe $SP_{A,M}$, die jedoch in der Stichprobe $SP_{B,M}$ eindeutig nicht signifikant wird. Zunächst suggerieren die resultierenden Mittelwertunterschiede, dass im Gegensatz zu den bisherigen Beobachtungen eine höhere Erwartungshaltung nicht zu einer positiveren Relevanzbewertung führt. Vielmehr bewerten Nutzer mit der niedrigeren Erwartung einen höheren Anteil der Dokumente richtig als eher relevant. Gleichzeitig akzeptieren jedoch Probanden mit der höheren Erwartungsmanipulation mehr eher irrelevante Dokumente als eher relevant (M31), wenngleich dieser Unterschied auch nicht signifikant ist. Somit erklärt diese weniger restriktive Bewertung eher irrelevanter Dokumente im Kontext der hohen Erwartungshaltung, warum der Anteil richtig als eher relevant bewerteter Dokumente für diese Testpersonen niedriger ausfällt. Zusammenfassend impliziert also eine höhere Erwartungshaltung wiederum eine positivere Relevanzbewertung.

Weitere Indizien dafür, dass sich der systembedingte Anpassungseffekt im Wesentlichen zwischen den beiden mittleren Relevanzkategorien abspielt, liefern die Ergebnisse für die Leistungsmaße V41, V42 und V43. Hierbei zeigt wiederum ausschließlich der Anteil der eher relevanten Dokumente, die fälschlicherweise als eher irrelevant bewertet werden (V41), einen signifikanten Systemeffekt. Die entsprechenden Leistungsmaße in Bezug auf relevante und irrelevante Dokumente (V42 u. V43) hingegen zeigen eindeutig bzw. in der Tendenz keine Systemabhängigkeit. Ergänzend zu den Effekten für die Leistungsmaße V41, V42 und V43 ist ein entsprechend umgekehrter Systemeffekt in Bezug auf die fälschlicherweise als eher relevant bewerteten Dokumente zu beobachten (M31, M32, M33, V45, V46, V47). Während wiederum kein Effekt für die beiden Extremfälle auftritt (M32, M33, V46, V47), bewerten Nutzer des besseren Systems signifikant weniger eher irrelevante Dokumente als eher relevant. Dieses Verhalten kann wiederum als systemabhängiger Anpassungseffekt der Relevanzwahrnehmung aufgefasst werden. In Bezug auf die Stabilität der Ergebnisse ist zu bemerken, dass bis auf M27, V41 und V45 alle Effekte in beiden Stichproben $SP_{A,M}$ und $SP_{B,M}$ mindestens in der Tendenz nachweisbar sind. Der Effekt in Bezug auf M27 hingegen ist nur für die Stichprobe $SP_{B,M}$ eindeutig signifikant, während der Effekt für $SP_{A,M}$ in drei Stichproben erkennbar ist. Für V45 kann die Systemabhängigkeit eindeutig nur für $SP_{A,M}$ nachgewiesen werden, während für $SP_{B,M}$ wiederum nur drei Stichproben signifikant werden. Das Leistungsmaß V41 schließlich weist nur bezüglich $SP_{B,M}$ einen Systemeffekt auf, der in $SP_{A,M}$ eindeutig nicht signifikant wird. Die genannten eindeutig nicht signifikanten Leistungsvariablen hingegen zeigen dieses Verhalten in Bezug auf beide Stichproben, lediglich das Leistungsmaß V46 wird ausschließlich in der Stichprobe $SP_{B,M}$ eindeutig nicht signifikant. Abbildung 7.13 stellt die beschriebenen Anpassungseffekte der Relevanzwahrnehmung in Bezug auf die mittleren Bewertungskategorien eher relevant und eher irrelevant noch einmal graphisch dar. Im Vergleich zu den Jurorenurteilen führt eine niedrige Systemleistung bzw. eine hohe Erwar-

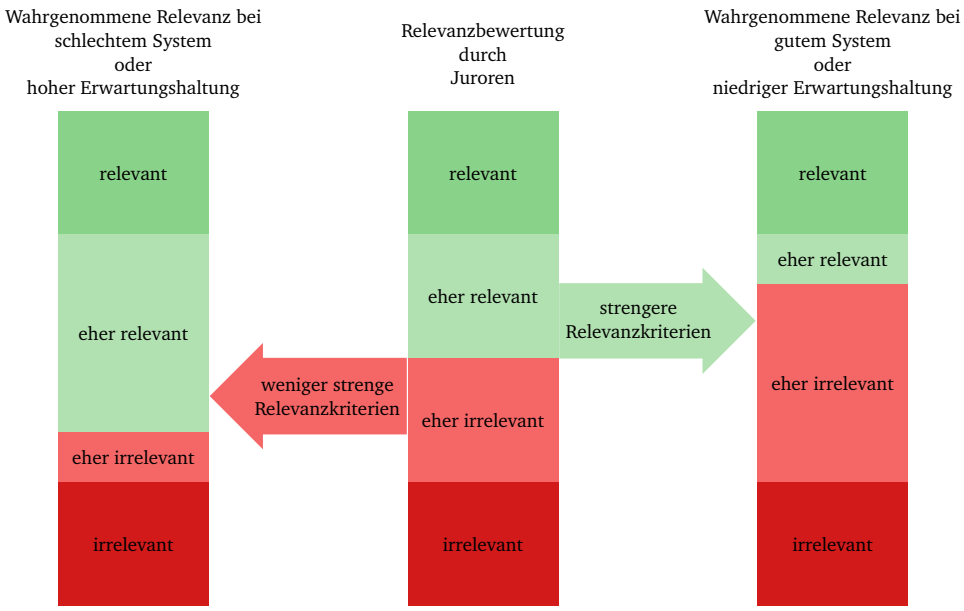


Abb. 7.13.: Schematische Darstellung der Änderung der Relevanzwahrnehmung in Bezug auf die mittleren Bewertungskategorien eher relevant und eher irrelevant.

tungshaltung zu einer weniger restriktiven Relevanzeinschätzung, wodurch mehr eher irrelevante Dokumente als eher relevant akzeptiert werden (rechtes Balkendiagramm). Im Fall des besseren Systems bzw. der geringeren Erwartungshaltung hingegen kommt es zu einer entgegengesetzten Verschiebung dieser Grenze (linkes Balkendiagramm). Die beiden extremen Relevanzkategorien relevant und irrelevant sind von diesem Effekt hingegen kaum betroffen.

Zusammenfassend lässt sich somit festhalten, dass die 4-stufige Relevanzskala in der Tat einen Mehrwert für die Evaluierung von IR-Systemen bietet, indem sie einen differenzierteren Blick auf die beobachteten Anpassungseffekte erlaubt. Im vorliegenden Fall kann damit gezeigt werden, dass die Bewertungsanpassungen sowohl in Bezug auf die Systemleistung, als auch für die Erwartungshaltung vorrangig im Bereich der mittleren Relevanzkategorien stattzufinden scheinen. Darüber hinaus kann für die Anpassung der Relevanzkriterien eine Wechselwirkung zwischen Systemgüte und Erwartungshaltung identifiziert werden, die in Bezug auf die binäre Bewertungsskala nicht sichtbar ist.

7.4.3.3. Dynamische Entwicklung der Benutzerleistung

Um Lerneffekten und anderen Prozessen, die sich erst aus der wiederholten Interaktion der Probanden mit dem Testsystem ergeben, Rechnung zu tragen, wird im Folgenden die Dynamik des Suchprozesses in die Auswertung einbezogen. Dazu werden die Daten des dritten Experiments noch einmal im Rahmen eines gemischten ANOVA-Designs ausgewertet. Diese Auswertungsmethode erlaubt neben den unabhängigen Variablen Systemleistung und Erwartungshaltung auch die Aufgabenposition als Messwiederholungsfaktor zu berücksichtigen (vgl. Abschn. 4.3.2.4). Da die Testpersonen jeweils drei Aufgaben bearbeiten, ergibt sich nun für den Untersuchungsplan ein $2\times2\times3$-Design (vgl. Abb. 7.1). Auf diese Weise können also auch Unterschiede in Bezug auf die wiederholte Interaktion der Probanden mit dem Testsystem durch den Innersubjektfaktor Aufgabenposition erfasst werden. Dies ermöglicht insbesondere die Untersuchung der Forschungshypothese H4, die besagt, dass Testpersonen im Laufe der Systemnutzung unrealistische

Erwartungen an die Systemgüte anpassen.

Von 114 Leistungsmaßen, für die eine Analyse aufgrund ausreichender Fallzahlen durchführbar ist, zeigen 41 Variablen signifikante Effekte. Dabei handelt es sich in 21 Fällen um einen signifikanten Einfluss der Aufgabenposition. Zu beachten ist weiterhin, dass für SP_B aufgrund der Stichprobengrößen ausschließlich Variablen ohne Topiceffekt in die Betrachtung eingehen ($SP_{B,OT}$). Für SP_A hingegen werden sowohl Variablen mit ($SP_{A,MT}$) als auch ohne Topiceffekt ($SP_{A,OT}$) analysiert. Wie in Abschnitt 7.4.2 beschrieben wird dazu für letzteren Fall die Topic-Reihenfolge über alle Aufgabenpositionen hinweg balanciert. Da im Rahmen der Auswertung abhängig vom Vorliegen oder der Abwesenheit eines Topiceffekts die jeweils notwendige Balancierungsstrategie gewählt wird, werden diese Qualitätstufen im Folgenden nicht mehr explizit im Text unterschieden sondern zusammenfassend als SP_A bzw. SP_B bezeichnet.

Die Ergebnisse der Varianzanalysen sind in den Tabellen 7.15 und 7.16 zusammengefasst. Um eine möglichst kompakte Präsentation der Ergebnisse zu ermöglichen, beschränkt sich die Darstellung in diesen Tabellen auf Variablen, die in mindestens vier der fünf getesteten Stichproben einen signifikanten Effekt der Aufgabenposition (Tab. 7.15) oder Interaktionen zwischen den drei Haupteffekten (Tab. 7.16) aufweisen. Da für die Stichprobe SP_B keine neuen Variablen mit signifikanter Aufgabenposition hinzukommen und die Ergebnisse bis auf das Signifikanzniveau einzelner Mittelwertsunterschiede übereinstimmen, werden hier ausschließlich die Ergebnisse der weniger kontrollierten Stichprobe SP_A berichtet. Die Ergebnisse der Stichprobe SP_B hingegen sowie alle weiteren signifikanten Haupteffekte, also solche, die ausschließlich Systemleistung und Erwartungshaltung betreffen, sind den Tabellen E.26 bis E.47 in Anhang E.4 zu entnehmen. Wie in den vorangegangenen Abschnitten beinhalten diese neben den Mittelwerten weitergehende Informationen, wie Teststatistiken, Stichprobengrößen und Signifikanzniveaus. Der Diskussion der dynamischen Entwicklung der Benutzerleistung ist jedoch zunächst ein Vergleich der dynamikunabhängigen Ergebnisse mit den in Abschnitt 7.4.3.1 beschriebenen Mittelwertanalysen vorangestellt.

Die wesentlichen in Abschnitt 7.4.3.1 berichteten Ergebnisse für die über alle drei Aufgaben gemittelten Leistungsmaße bleiben unter der Einbeziehung des Messwiederholungsfaktors Aufgabenposition stabil. Dies gilt insbesondere für die Anpassung der Relevanzwahrnehmung in Abhängigkeit von Systemleistung und Erwartungshaltung. So zeigt sich im Kontext der Systemleistung wiederum, dass Benutzer des besseren Systems zu einer restriktiveren Relevanzbewertung neigen bzw. die Zustimmungstendenz zu den Jurorenurteilen mit steigender Systemleistung abnimmt. Konkret kann für acht der elf Benutzerleistungsmaße, für die in Abschnitt 7.4.3.1 ein über $SP_{A,M}$ und $SP_{B,M}$ stabiler Systemanpassungseffekt nachgewiesen wird, die dynamische Auswertung aufgrund ausreichender Stichprobengrößen durchgeführt werden (M07, M14, M15, V05, V06, V13, V14 u. V17). In allen Fällen wird der Effekt in SP_A bestätigt. In fünf Fällen (M07, V05, V06, V14 u. V17) kann er sogar in SP_A und in SP_B nachgewiesen werden, ist er also erneut bzgl. beider Stichproben stabil.

Darüber hinaus ist auch der erwartungsinduzierte Anpassungseffekt im dynamischen Kontext weiterhin sichtbar. Dabei bestätigen sich die Befunde für die Leistungsmaße B04 und B06 (SP_B) sowie V11 (SP_A u. SP_B) und V54 (SP_A). Allerdings fällt der Effekt nicht mehr so deutlich aus, da im Vergleich zur gemittelten Auswertung die Mittelwertunterschiede für die Variablen B18, V08,

V12 und V14 zwar in die richtige Richtung weisen, aber nicht länger signifikant sind. Für die Leistungsmaße B04 und B06 lässt sich weiterhin der Wechsel von einem Erwartungsanpassungseffekt in SP_B zu einem Systemanpassungseffekt in SP_A beobachten.

Im Folgenden werden die Ergebnisse für den Haupteffekt Aufgabenposition berichtet. Die entsprechenden Mittelwerte sind in der Tabelle 7.15 enthalten. Wie eingangs erwähnt, werden an dieser Stelle ausschließlich die Ergebnisse der Stichprobe SP_A berichtet, da in SP_B keine neuen Effekte der Aufgabenposition hinzukommen. Die entsprechenden Ergebnisse für SP_B sind jedoch in Tabelle E.26 in Anhang E.4 aufgeführt. Wie bei der Analyse der gemittelten Leistungsmaße sind für jeden signifikanten Haupteffekt jeweils die Mittelwerte der Stichprobe mit dem niedrigsten p-Wert angegeben. Wiederum sind Haupteffekte, deren Einfluss auf die abhängige Variable signifikant ist, in der Tabelle fett hervorgehoben. Für den dreistufigen Faktor Aufgabenposition ist es darüber hinaus erforderlich, Post-hoc-Tests durchzuführen, um die tatsächlich signifikanten Mittelwertunterschiede zwischen den drei Positionen zu ermitteln. Diese sind in der Tabelle durch die Symbole $<$ und $>$ markiert, die gleichzeitig die Richtung des Effekts angeben. Die Spalte hinter Position 3 bezieht sich dabei auf den Unterschied zwischen der zuerst und zuletzt bearbeiteten Suchaufgabe. Darüber hinaus weisen Fußnoten in der Tabelle die Treatmentgruppe mit der besseren Nutzerleistung sowie eventuell aufgetretene numerische Instabilitäten bei der Berechnung der Post-hoc-Tests aus.

Eine natürliche Herangehensweise, um die dynamische Entwicklung der Benutzerleistung zu analysieren, besteht darin zunächst Lern- und Ermüdungseffekte zu identifizieren. In der Tat lassen sich viele der vorliegenden Aufgabenpositionseffekte einer dieser beiden Gruppen zuordnen. Dabei werden die folgenden Definitionen zugrunde gelegt: Ein Lerneffekt liegt vor, sobald die dritte Aufgabe signifikant besser als die ersten beiden bearbeitet wird, also sowohl der Mittelwertunterschied zwischen Position 1 und 3 als auch zwischen 2 und 3 signifikant ist. Gleiches gilt für den Fall, dass die Ergebnisse der zweiten und dritten Aufgabe im Vergleich zur ersten Aufgabe beide signifikant besser ausfallen. Ermüdungseffekte sind hingegen durch ein gegenteiliges Verhalten gekennzeichnet. Hier müssen entweder die letzten beiden Aufgaben signifikant schlechter als die erste oder die ersten beiden Aufgaben signifikant besser als die letzte Aufgabe bearbeitet werden.

Der auffälligste Effekt des Messwiederholungsfaktors Aufgabenposition ergibt sich hinsichtlich der durchschnittlichen Betrachtungszeiten der von den Probanden aufgerufenen Dokumente. In dreizehn Fällen lässt sich diesbezüglich eine signifikante Abnahme der Betrachtungsdauer nachweisen (S04, S05/S05-log, Z01/Z01-log, Z05, Z07-log, Z08/Z08-log, Z09/Z09-log, Z11/Z11-log, Z14/Z14-log, Z15/Z15-log, Z22/Z22-log, Z23/Z23-log, Z28-log). Von der ersten zur letzten Aufgabenposition verringern sich die Betrachtungszeiten dabei um 11 bis 23 Sekunden, was einem prozentualen Anteil von 25 % bis 33 % entspricht. Ein ähnliches Verhalten ist für den Übergang von Aufgabenposition 1 zu 2 zu beobachten, wo sich die Betrachtungsdauer um $7s$ bis $19s$ verkürzt, was einer Reduktion um 15 % bzw. 30 % entspricht. Insgesamt gesehen fällt dieser Effekt also relativ deutlich aus. Des Weiteren geht auch die Bearbeitungszeit der Suchaufgaben (S04) um gut $60s$ zurück. Die Gesamtheit dieser Beobachtungen lässt zwei offensichtliche Interpretationen zu. Zum einen könnte die Reduktion der Zeiten darauf hinweisen, dass die Probanden nach der initialen Aufgabe besser mit dem System zurechtkommen und deshalb im Folgenden ihre

Tab. 7.15.: Signifikante Positionseffekte der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Größer-/Kleinerzeichen ($\gtrless$) zwischen den Positionsmittelwerten (P_i) markieren signifikante Mittelwertsunterschiede. Dabei bezieht sich die letzte Spalte auf den Vergleich zwischen P_3 und P_1. Informationen zu Interaktionen und p-Werten können Tabelle E.40 in Anhang E.4 entnommen werden.

ID	Beschreibung	S_G	S_S	E_H	E_N	P_1	P_2	P_3
M01	Anz. aufg. Dok.	8,41	8,71	8,64	8,49	**7,32** <	**9,31**[b] >	**9,06** >
M06[a]	Anz. aufg. rel. Dok.	6,50	5,92	6,55	5,87	**5,37** <	**6,85**[b] >	**6,40** >
M07[a]	Anz. falsch irrel. bew. Dok.	**2,34** >	**1,61**[b] >	1,97	1,99	**1,77** <	**2,50** >	**1,66**[b] <
M10	Anz. rel. bew. Dok.	4,93	5,21	5,39	4,75	**4,38** <	**5,28**	**5,55**[b] >
M20[a]	Anz. aufg. eher irrel. Dok. (4-st.)	**0,90**[b] <	**1,66** >	1,22	1,33	**0,93**[b] <	**1,41**	**1,50** >
M26[a]	Anz. falsch eher irrel. bew. Dok. (4-st.)	1,49	1,22	1,32	1,39	**1,16** <	**1,76** >	**1,14**[b] <
M37[a]	Anz. rel. bew. Dok. (4-st.)	2,60	2,84	2,87	2,57	**2,20** <	**2,93**	**3,03**[b] >
Z01	Durchschn. Betrachtungsz. aller Dok.	52,54	51,60	53,87	50,28	**62,54** >	**48,38**	**45,30**[b] <
Z01-log	Durchschn. Betrachtungsz. aller Dok.	3,60	3,54	3,61	3,53	**3,77** >	**3,53** >	**3,42**[b] <
Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	37,57	34,45	37,40	34,62	**42,32** >	**35,04** >	**30,68**[b] <
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	3,77	3,78	3,75	3,80	**3,98** >	**3,75** >	**3,60**[b] <
Z08	Durchschn. Betrachtungsz. rel. Dok.	56,69	54,65	58,15	53,20	**67,61** >	**51,67** >	**47,73**[b] <
Z08-log	Durchschn. Betrachtungsz. rel. Dok.	3,71	3,70	3,68	3,73	**3,88** >	**3,68** >	**3,56**[b] <
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	54,30	57,03	58,02	53,32	**69,41** >	**50,72**	**46,88**[b] <
Z09-log	Durchschn. Betrachtungsz. richtig bew. Dok.	3,77	3,47	3,60	3,64	**3,85** >	**3,58**	**3,44**[b] <
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	59,89	59,60	58,73	60,76	**71,91** >	**57,19** >	**50,14**[b] <
Z11-log	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	3,78	3,72	3,74	3,75	**3,96** >	**3,72** >	**3,57**[b] <
Z14-log	Durchschn. Betrachtungsz. eher rel. bew. Dok. (4-st.)	3,72	3,72	3,67	3,76	**3,91**	**3,74** >	**3,51**[b] <
Z15	Durchschn. Betrachtungsz. eher rel. Dok. (4-st.)	50,07	43,13	49,08	44,11	**59,14** >	**41,48**	**39,17**[b] <
Z23	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	56,62	57,23	56,51	57,35	**65,30** >	**55,19**	**50,29**[b] <
Z23-log	Durchschn. Betrachtungsz. rel. Dok.	3,72	3,62	3,60	3,73	**3,85** >	**3,67** >	**3,48**[b] <
V03[a]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	0,08	0,08	0,08	0,08	**0,07** <	**0,09**[b]	**0,08** >
V04	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	0,12	0,13	0,12	0,12	**0,10** <	**0,13**[b]	**0,12** >
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,30**	**0,19**[b]	0,24	0,25	**0,26**	**0,27** >	**0,20**[b] <
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	0,68	0,75	0,72	0,71	**0,71**	**0,67** <	**0,76**[b] >
V32/BR	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	0,06	0,06	0,06	0,06	**0,05** <	**0,06**	**0,07**[b] >
V33	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	0,08	0,09	0,09	0,08	**0,08** <	**0,09**	**0,09**[b] >
V38[a]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$ (4-st.)	0,10	0,10	0,11	0,09	**0,08** <	**0,11**[b]	**0,10** >
S04	Suchdauer	489,42	486,70	488,67	487,45	**507,12**	**512,16** >	**444,90** <
S05-log	Zeit zum ersten richtig rel. bew. Dok.	4,85	4,81	4,83	4,82	**5,04** >	**4,82** >	**4,61**[b] <

[a] Warnung bei Posthoc-Test der Positionsmittelwerte wegen numerischer Instabilität.
[b] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Relevanzentscheidung effizienter treffen können. Zum anderen jedoch könnten die verkürzten Betrachtungszeiten auf Ermüdungseffekte der Testteilnehmer hindeuten. Dieser letzten Interpretation widerspricht jedoch die Tatsache, dass gleichzeitig die Zahl der aufgerufenen Dokumente

(M01) mit der Aufgabenposition zunimmt. Weiterhin verkürzt sich die Zeit, die die Probanden benötigen, um das erste richtig als relevant bewertete Dokument zu finden (S05). Beide Effekte lassen sich als Hinweise darauf interpretieren, dass die Verkürzung der Betrachtungszeiten als Lerneffekt der Probanden gewertet werden kann.

Auch der Großteil der übrigen signifikanten Effekte der Aufgabenposition weist auf Lerneffekte der Teilnehmer hin (M01, M06, M10, M23, M37, M40, V04, V05, V17, V32/BR u. V33). So rufen sie im Verlauf des Tests bspw. nicht nur insgesamt mehr Dokumente auf (M01), sondern es erhöht sich zudem insbesondere die Anzahl der aufgerufenen relevanten Dokumente unabhängig davon, ob die binäre (M10) oder 4-stufige Relevanzskala (M23) zugrunde gelegt wird. Weiterhin steigt im binären Fall die Anzahl der relevant bewerteten Dokumente (M06). Bemerkenswerterweise geht dies mit einer Zunahme verschiedener Recallmaße (V04, V32/BR u. V33) einher, die im Kontext der System- und Erwartungsabhängigkeit im Gegensatz zu den Precisionmaßen noch keine signifikanten Effekte gezeigt haben. So nehmen sowohl der Anteil der aufgerufenen relevanten Dokumente (V04) als auch der Anteil der richtig als relevant bewerteten Dokumente (V32/BR u. V33) zu. Gleichzeitig sinkt der Anteil der fälschlicherweise als irrelevant identifizierten relevanten Dokumente (V05 u. V17). Für das Leistungsmaß M20 lässt sich ein Ermüdungseffekt beobachten. So ist der Anstieg bei den aufgerufenen Dokumenten (M01) nicht allein auf eine höhere Anzahl aufgerufener relevanter Dokumente (M06) zurückzuführen, sondern wird in Teilen durch eine Zunahme der aufgerufenen eher irrelevanten Dokumente verursacht (M20). Die Probanden scheinen im Testverlauf also auch in weniger relevanten Dokumenten nach Informationen zu suchen.

Über die drei Haupteffekte hinaus wird mit Ausnahme von M27 in SP_B zunächst keine der vier möglichen Wechselwirkungen eindeutig oder in der Tendenz signifikant. Dies ist insbesondere darauf zurückzuführen, dass nur wenige Stichproben die Voraussetzungen für eine klassische Varianzanalyse erfüllen und somit die robuste Variante mit geringerer statistischer Stärke gewählt wird. Da jedoch bekannt ist, dass sich das klassische Varianzanalyseverfahren relativ robust in Bezug auf eine Verletzung ihrer Voraussetzungen verhält (Bortz, 2005, S. 352), werden im Folgenden die Ergebnisse der klassischen Auswertung mit signifikanten Wechselwirkungen diskutiert. Dieses Vorgehen erlaubt eine weitergehende Analyse der Dynamik der Benutzerleistung, wenngleich eine Generalisierbarkeit der Ergebnisse in diesem Fall nur eingeschränkt möglich ist.

Bei Betrachtung der klassischen Varianzanalysen werden sowohl für SP_A als auch für SP_B bei jeweils vier Leistungsmaßen Wechselwirkungen zwischen den Faktoren eindeutig oder in der Tendenz signifikant. Dabei handelt es sich um die Variablen B05, S04 und Z07 (SP_A) sowie V11, V12 und V28/PCP (SP_B). Darüber hinaus ist das Leistungsmaß M27 stabil über beide Stichproben hinweg eindeutig signifikant. Die entsprechenden Ergebnisse sind in Tabelle 7.16 zusammengefasst. Im Folgenden werden die beobachteten Effekte getrennt nach den entsprechenden Wechselwirkungen dargestellt.

Signifikante Wechselwirkungen zwischen Systemleistung und Erwartungshaltung sind ausschließlich für die durchschnittliche Betrachtungsdauer relevant bewerteter Dokumente bei binärer Relevanzskala zu beobachten (Z07). Wie Abbildung 7.14 Bild (a) zu entnehmen ist, führen die Kombinationen hohe Erwartungshaltung/schlechtes System und niedrige Erwartungshaltung/gutes System jeweils zu einer längeren Betrachtungsdauer der Dokumente. Eine Diskrepanz

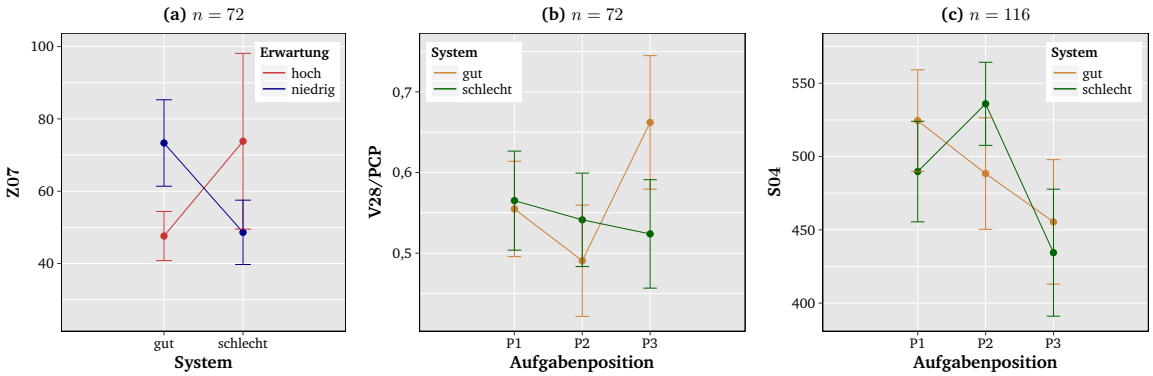


Abb. 7.14.: Wechselwirkungen im Rahmen der klassischen gemischten Varianzanalyse für die Leistungsmaße Z07, V28/PCP und S04. Bild (a): Signifikante Wechselwirkung zwischen Systemleistung und Erwartungshaltung für die durchschnittlich Betrachtungsdauer relevant bewerteter Dokumente (Z07). Entspricht die Erwartungshaltung nicht der präsentierten Systemleistung, ist jeweils eine Zunahme der Betrachtungszeit zu beobachten. Bild (b): Signifikante Wechselwirkung zwischen Systemgüte und Aufgabenposition für den Anteil der richtig als relevant bewerteten Dokumente an allen aufgerufenen Dokumenten (V28/PCP). Die im Rahmen der Mittelwertsauswertung beobachtete höhere Precision im Kontext des besseren Systems wird von den Teilnehmern erst für die letzte Aufgabe erreicht. Bild (c): Signifikante Wechselwirkung zwischen Systemgüte und Aufgabenposition für die Aufgabenbearbeitungszeit (S04). Während Nutzer des besseren Systems tendenziell ihre Suchdauer im Laufe des Tests kontinuierlich reduzieren zu scheinen, benötigen die Nutzer des schlechteren Systems zur Lösung der zweiten Aufgabe signifikant länger. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

zwischen Erwartungshaltung und Systemgüte scheint somit direkt zu einer höheren Betrachtungszeit zu führen, unabhängig davon, ob die Erwartung der Probanden enttäuscht oder übertroffen wird.

Mit fünf signifikanten Leistungsmaßen (V11, V12, V28/PCP, B05 u. S04) ist die Interaktion zwischen Systemleistung und Aufgabenposition am stärksten ausgeprägt. In drei Fällen (V11, V12 u. V28/PCP) handelt es sich dabei um Precisionmaße, bei denen in Bezug auf die Mittelwertanalyse zu beobachten ist, dass eine höhere Systemgüte generell zu einer besseren Benutzerleistung führt. Die signifikante Wechselwirkung mit der Aufgabenposition präzisiert diesen Effekt als einen Lerneffekt: Die bessere Leistung erreichen die Probanden mit der höheren Systemgüte erst bei der letzten Aufgabe. Für die ersten beiden Aufgaben hingegen lässt sich noch kein Unterschied in Bezug auf die Systemgüte feststellen. Beispielhaft ist dieses Verhalten für V28/PCP in Abbildung 7.14 Bild (b) dargestellt.

Die signifikante Wechselwirkung in Bezug auf die Suchdauer S04, deren Abhängigkeit von Aufgabenposition und Systemleistung in Abbildung 7.14 Bild (c) dargestellt ist, deutet auf unterschiedliche Lernkurven für die beiden Systemqualitäten hin. Während die Suchdauer und damit der Aufwand, den die Benutzer zur Lösung ihrer Suchaufgaben betreiben müssen, für das bessere System kontinuierlich über alle drei Aufgaben abnimmt, steigt die Bearbeitungszeit in der Testgruppe mit der geringeren Systemleistung zunächst an. Zur Bearbeitung der letzten Aufgabe hingegen benötigen beide Treatmentgruppen in etwa wieder gleich lang. Hier wäre

Tab. 7.16.: Signifikante Interaktionseffekte der Varianzanalysen zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung bei binärer und 4-stufiger Relevanzskala in SP_A und SP_B. Dargestellt sind die Ergebnisse des klassischen Analyseverfahrens auch bei Verletzung der statistischen Voraussetzungen. Angegeben sind in mindestens in vier von fünf Stichproben nachweisbare Interaktionen zwischen Systemgüte (S), Erwartungshaltung (E) und Position (P). I_{SEP} bezeichnet bspw. eine dreifache Wechselwirkung zwischen Systemgüte, Erwartungshaltung und Position. p-Werte und weitergehende Informationen über Gruppengröße und Teststatistik können den Tabellen E.46 und E.47 in Anhang E.4 entnommen werden.

		ID	Beschreibung	sig. Interaktion
SP_A	binär	B05	Durchschn. Bew. rel. Dok. (erste Suche)	I_{SP}
		Z07	Durchschn. Betrachtungsz. rel. bew. Dok.	I_{SE}
		S04	Suchdauer	I_{SP}
	4-st.	M27[a]	Anz. falsch eher irrel. bew. eher rel. Dok.	I_{SEP}
SP_B	binär	V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	I_{SP}
		V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	I_{SP}
		V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	I_{SP}
	4-st.	M27[a]	Anz. falsch eher irrel. bew. eher rel. Dok.	I_{SEP}

[a] Diese Wechselwirkung ist in der Tendenz auch für die robuste Analyse in SP_B nachweisbar.

es sicherlich interessant die Bearbeitungszeit über mehr als drei Aufgaben zu verfolgen, um zu sehen, ob der relativ deutliche Unterschied in der Suchdauer von ca. $40s$ tatsächlich nur nach der zweiten Aufgabe zu beobachten ist.

Besonders bemerkenswert ist das Verhalten des Leistungsmaßes B05, d.h. der durchschnittlichen Bewertung relevanter Dokumente bei der ersten Suche in Bezug auf die binäre Relevanzskala. Wie Bild (a) in Abbildung 7.15 zu entnehmen, schätzen Nutzer des schlechteren Systems bei den ersten beiden Aufgaben die Relevanz relevanter Dokumente höher ein als Nutzer des besseren Systems. Dieses Verhalten spiegelt also einmal mehr den bereits beschriebenen systembedingten Anpassungseffekt der Relevanzwahrnehmung wider. Bei der dritten Aufgabe hingegen dreht sich dieses Verhalten um und Nutzer des besseren Systems zeigen nun eine positivere Relevanzwahrnehmung im Vergleich zu der Testgruppe mit dem schlechteren System. Dieses Verhalten widerspricht der Vermutung, dass der Systemanpassungseffekt durch die fortgesetzte Interaktion mit dem System stabilisiert wird und sich verfestigt. Um diesen Widerspruch aufzuklären wird in Abbildung 7.15 Bild (b) das zeitliche Verhalten der durchschnittlichen Bewertung relevanter Dokumente bei der ersten Suche (B05) der durchschnittlichen Bewertung relevanter Dokumente bei der letzten Suche (B06) gegenübergestellt. Bemerkenswerterweise zeigt B06 genau das erwartete Verhalten eines durchgehend vorhandenen signifikanten Systemanpassungseffekts, bei dem Nutzer des schlechteren Systems durchgehend zu einer positiveren Bewertung der relevanten Dokumente kommen. Zusätzlich zu der zeitlichen Abhängigkeit über die einzelnen Aufgabenpositionen hinweg zeigt sich somit noch eine weitere dynamische Komponente, die vom Bewertungszeitpunkt innerhalb der einzelnen Suchaufgaben abhängt: Während Leis-

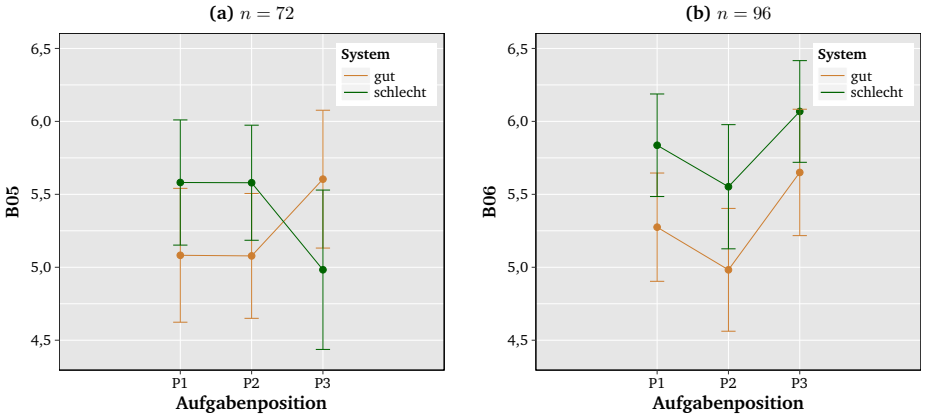


Abb. 7.15.: Vergleich des dynamischen Verhaltens der durchschnittlichen Bewertung relevanter Dokumente im Rahmen der ersten (B05) und der letzten Suche (B06). Während die signifikante Wechselwirkung zwischen Systemleistung und Aufgabenposition auf eine Umkehrung des systembedingten Anpassungseffekts der Relevanzwahrnehmung hinzuweisen scheint, bei der bei der letzten Aufgabe Nutzer des besseren Systems zu einer positiveren Einschätzung der Relevanz kommen (a), zeigt sich für B06 hingegen durchgängig eine strengere Relevanzbewertung im Kontext der besseren Systemleistung (b). Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

tungsmaß B05 auf der ersten Bewertung der Dokumente beruht, geht in die Berechnung von B06 ausschließlich die letzte Bewertung eines Dokuments ein.

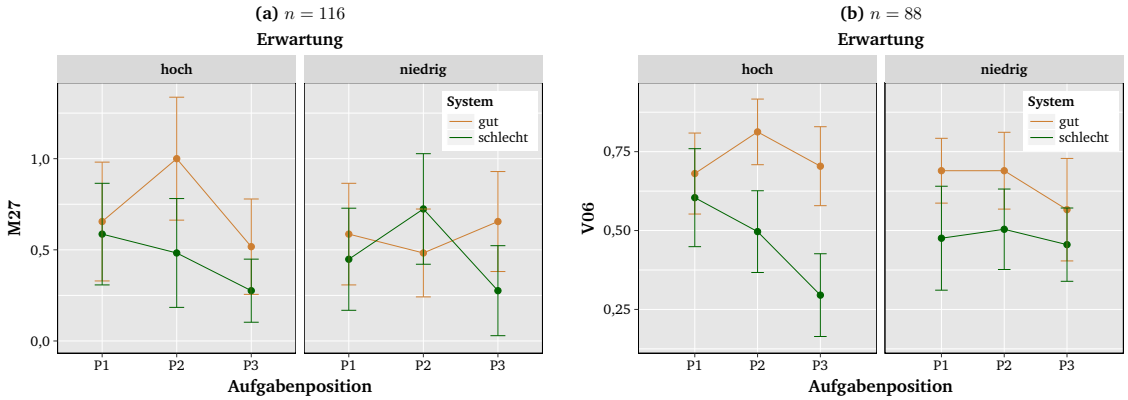


Abb. 7.16.: Wechselwirkung zwischen Systemleistung, Erwartungshaltung und Aufgabenposition für die Anzahl der eher relevanten Dokumente, die fälschlicherweise als irrelevant bewertet werden (M27) sowie den Anteil fälschlicherweise als irrelevant bewerteter Dokumente bei binärer Relevanzskala (V06). Während im Kontext der niedrigen Erwartungshaltung der Unterschied zwischen den beiden Systemleistungen im Wesentlichen konstant zu bleiben scheint, zeigt sich im Rahmen der hohen Erwartungshaltung eine dynamische Abhängigkeit, die sich im Fall von V06 durch eine stetige Abnahme für das schlechtere System äußert. Fehlerbalken kennzeichnen 95 %-Konfidenzintervalle der Gruppenmittelwerte.

Mit M27 wird für ein Effektivitätsmaß über beide Stichproben SP_A und SP_B hinweg eine Dreifachwechselwirkung zwischen Systemgüte, Erwartungshaltung und Aufgabenposition signifikant. Es handelt sich dabei um die Anzahl der eher relevanten Dokumente, die bei 4-stufiger Relevanz-

skala falsch als eher irrelevant bewertet werden und somit um ein Leistungsmaß, dass eng mit dem Systemanpassungseffekt verbunden ist. Der entsprechenden Interaktionsgraph ist in Abbildung 7.16 Bild (a) dargestellt. Im Falle der höheren Erwartungshaltung führt eine bessere Systemleistung erneut zu einer strengeren Relevanzbewertung der Testteilnehmer: Probanden mit dem besseren System bewerten mehr eher relevante Dokumente fälschlicherweise als eher irrelevant. Darüber hinaus zeigt sich jedoch, dass dieses Verhalten an der ersten Aufgabenposition noch nicht erkennbar ist. Der Anpassungseffekt scheint also Zeit bzw. Kontakt mit dem Suchsystem zu benötigen um sich auszubilden. Dies gilt in noch stärkerem Maße für Teilnehmer mit der niedrigen Erwartungshaltung. Hier ist der Systemanpassungseffekt erst an der dritten Aufgabenposition erkennbar, während die Unterschiede in Bezug auf die Systemgüte für die ersten beiden Aufgabenpositionen nur einen geringen Unterschied aufweisen (vgl. Abb. 7.16 (a)). Im Fall der hohen Erwartungshaltung scheinen die Probanden darüber hinaus im Fall des schlechteren Systems über die drei Aufgabenpositionen hinweg tendenziell immer positiver zu urteilen, da der Wert von M27 kontinuierlich abnimmt. Im Kontext des besseren Systems stellt sich das Verhalten jedoch komplizierter dar, da sich die Relevanzeinschätzung an Aufgabenposition drei nach einer sehr strengen Bewertung im Rahmen der zweiten Aufgabe wieder in etwa auf dem Niveau von Aufgabe eins befindet. Ein ähnliches Verhalten lässt sich des Weiteren für das Imprecisionmaß V06 beobachten, das den Anteil der fälschlicherweise als irrelevant bewerteten Dokumente bei binärer Relevanzskala erfasst. Allerdings wird V06 nur in drei Stichproben von SP_A signifikant. Nichtsdestotrotz ist es interessant zu sehen, dass auch hier die Erwartungshaltung ein unterschiedliches dynamisches Verhalten des Systemanpassungseffekts induziert. Während der Effekt im Falle der niedrigen Erwartungshaltung im Wesentlichen über die Aufgabenpositionen konstant bleibt, führt eine hohe Erwartungshaltung dazu, dass der Unterschied zwischen hohem und niedrigem System über die Zeit zunimmt.

Zusammenfassend lässt sich zunächst festhalten, dass die in Rahmen der Mittelwertanalysen erhaltenen Ergebnisse unter der Berücksichtigung des dynamischen Nutzerverhaltens stabil weiterhin nachweisbar sind, was insbesondere auf die beiden beschriebenen Anpassungseffekte zutrifft. Darüber hinaus ermöglicht die hier durchgeführte, dynamische Analyse, ähnlich der erweiterten 4-stufigen Relevanzskala, einen detaillierteren Einblick in die dynamische Abhängigkeit dieser Anpassungseffekte von Systemleistung und Erwartungshaltung.

7.4.4. Skalenbildung

Genau wie im zweiten Experiment geht der Auswertung der Benutzerzufriedenheit wiederum eine explorative Faktorenanalyse voraus. Der Prozess der Skalenbildung orientiert sich dabei an der in Abschnitt 6.4.4 diskutierten Vorgehensweise (vgl. Abb. 6.15). Entsprechend ist der vorliegende Abschnitt in drei Teile untergliedert. Zunächst wird eine Itemanalyse durchgeführt, um weniger geeignete Frageitems im Vorfeld identifizieren zu können. Im Anschluss erfolgt die explorative Faktorenanalyse mit den sich als geeignet erweisenden Items. Zusätzlich wird im letzten Schritt die Reliabilität und die Validität der Skalen anhand unterschiedlicher Gruppenvergleiche bewertet.

Die Datengrundlage bildet der unbalancierte, über alle drei Suchaufgaben gemittelte Datensatz $SP_{A,M}$ ohne fehlende Werte ($n = 128$), wobei unbalanciert bedeutet, dass pro Versuchsgruppe alle verfügbaren Fälle in die Skalenbildung einbezogen werden (vgl. Abschn. 7.4.2). Anders als

im zweiten Experiment geht somit jeder Teilnehmer nur einmal in die Analyse ein. Das Mittelungsverfahren bietet darüber hinaus den Vorteil, dass Schwankungen in den Einzelmessungen ausgeglichen werden.

7.4.4.1. Itemanalyse

Das Vorgehen zur Itemanalyse erfolgt wie in Abschnitt 6.4.4.1 beschrieben. Um der Frage nachzugehen, ob die verwendeten Frageitems das Zielmerkmal Benutzerzufriedenheit hinreichend genau definieren, werden wiederum zwei Korrelationsmaße berechnet. Diese quantifizieren wie gut das gesamte Testergebnis aufgrund der Beantwortung eines einzelnen Items vorhersagbar ist (Korrigierte Item-Total-Korrelation) und inwieweit dieses Item in erwarteter Weise mit der Messung eines externen Kriteriums übereinstimmt (Korrelation mit dem Kriterium).

Tab. 7.17.: Trennschärfe und Kriteriumsvalidität aller Zufriedenheitsitems ($n = 128$).

Item	Beschreibung	Korrigierte Item-Total-Korrelation	Korrelation mit dem Kriterium
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	0,83	0,82
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	0,85	0,82
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	0,82	0,78
F04	Liefert die Suchmaschine genügend Information?	0,83	0,81
F05	Ist die Suchmaschine präzise?	0,86	0,83
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	0,87	0,87
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,78	0,75
F08	Ist die Suchmaschine benutzerfreundlich?	0,62	0,59
F09	Ist die Suchmaschine einfach zu bedienen?	0,57	0,51
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	0,74	0,74
F11	Liefert die Suchmaschine aktuelle Information?	0,55	0,57
F14	Es war einfach, die Aufgabe zu bearbeiten.	0,66	0,63
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	0,85	0,82
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	0,71	0,67
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	0,83	0,78
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	0,82	0,78
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,67	0,60
F23	Ich bin mit meiner Suchleistung zufrieden.	0,72	0,72
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	0,70	0,66

Tabelle 7.17 zeigt die Ergebnisse der Itemanalyse aller Zufriedenheitsitems. Die Ergebnisse der zusätzlich separat durchgeführten Trennschärfeberechnungen bezüglich der EUCS- und Zusatzitems unterscheiden sich kaum von den Werten der Gesamtauswertung und können daher in Anhang E.1 nachvollzogen werden. Die geforderten Mindestwerte von $0{,}5$ für die Trennschärfe und $0{,}4$ für die Kriteriumsvalidität werden in den meisten Fällen deutlich überschritten und die geforderten Gütekriterien sind somit in allen Fällen erfüllt. Aus diesem Grund muss im dritten Experiment keines der Items von der Skalenbildung ausgeschlossen werden.

7.4.4.2. Explorative Faktorenanalyse

Nachdem die Analyse der Einzelitems abgeschlossen und die Eignung aller Items für die Faktorenanalyse belegt ist, wird im Folgenden die Durchführung selbiger beschrieben. Wie im zweiten Experiment wird dabei zunächst faktoranalytisch geprüft, ob sich die Struktur des EUCS-Instruments bestätigen lässt. Anschließend erfolgt zunächst eine separate Analyse der Zusatzitems, bevor abschließend alle Frageitems gesammelt in die Betrachtung einbezogen werden. Ziel dieser Analysen ist es, die im zweiten Experiment gefundene Faktorstruktur einer erneuten Prüfung zu unterziehen.

Tab. 7.18.: Eignung der Daten für eine explorative Faktorenanalyse.

Datensatz	#Items	Kaiser-Meyer-Olkin-Koeffizient $> 0{,}5$	Bartlett-Test auf Sphärizität $\leq 0{,}05$	Determinante der Korrelationsmatrix $> 1 \cdot 10^{-5}$
EUCS-Items	11	0,926 ausgezeichnet (superb)	$\chi^2(55) = 1212$ $p < 0{,}001$	$5{\cdot}10^{-5}$
nur Zusatzitems	8	0,909 ausgezeichnet (superb)	$\chi^2(28) = 696$ $p < 0{,}001$	0.004
alle Items	15	0,951 ausgezeichnet (superb)	$\chi^2(105) = 1696$ $p < 0{,}001$	$8{\cdot}10^{-7}$

Die Voraussetzungsprüfung zur Durchführung der Hauptkomponentenanalyse ist in Tabelle 7.18 zusammengefasst. Da das vorrangige Ziel, wie soeben beschrieben, in der Prüfung der Stabilität der im zweiten Experiment ermittelten Faktorstruktur besteht, addieren sich die Itemanzahlen in der letzten Zeile nicht auf 19 auf. Wie in der ersten Analyse werden die Items F06, F09, F10 und F11 von der Auswertung aller Zufriedenheitsitems ausgeschlossen, um einen vergleichbaren Datensatz zu erhalten. Für die EUCS-Items können die Voraussetzungen als ausgezeichnet angesehen werden (KMO-Kriterium: $0{,}93$; Bartlett-Test: $p < 0{,}001$). Multikollinearität scheint ebenfalls nicht vorzuliegen, da die Determinante der Korrelationsmatrix einen Wert knapp über der kritischen Grenze ($1 \cdot 10^{-5}$) besitzt. Um zu überprüfen, ob sich die ursprüngliche Faktorstruktur des EUCS-Instruments anhand der Daten des dritten Experiments reproduzieren lässt, wird zunächst eine Hauptkomponentenanalyse mit orthogonaler Varimax-Rotation für 5 Faktoren durchgeführt (vgl. Anh. E.2.1, Tab. E.3). Jedoch legen wie schon im Fall des zweiten Experiments auftretende Doppel- und Mehrfachladungen einzelner Items eine Korreliertheit der Faktoren nahe, sodass im Folgenden eine oblique Rotationsmethode gewählt wird. Da somit eine vollständige Replikation der von Doll und Torkzadeh (1988) beschriebenen Faktorstruktur nicht möglich ist, wird zunächst versucht, die im zweiten Experiment gewählte Dreifaktorenlösung zu replizieren. Da auch dies zu keiner gut interpretierbaren Faktorlösung führt, werden verschiedene Verfahren zur Bestimmung der Faktorenzahl herangezogen (Luhmann, 2013, S. 290 ff.). Der detailierte Verlauf der Analyse der EUCS-Items kann in Anhang E.2.1 nachvollzogen werden. Bis auf Parallelanalyse und Minimal Average Partial-Kriterium, die die Extraktion von nur einem Faktor nahelegen, sprechen die übrigen Verfahren (Eigenwertkriterium nach Jolliffe, Scree-Plot u. Very Simple Structure-Wert (VSS)) für eine Zweifaktorenlösung.

Tabelle 7.19 zeigt die rotierte Ladungsmatrix dieser Lösung nach der Oblimin-Rotation (KMO-Kriterium: $0{,}92$; Bartlett-Test: $p < 0{,}001$). Die beiden extrahierten Komponenten klären $80{,}4\,\%$ der Varianz auf. Auf der ersten Komponente laden die Items zu Inhalt, Genauigkeit und Darstellung der Suchergebnisse, auf der zweiten die Items zur allgemeinen Benutzerfreundlichkeit der Suchmaschine. Die internen Konsistenzen dieser Skalen ergeben zufriedenstellende Werte von $\alpha \geq 0{,}77$. Während also die ersten drei Skalen des EUCS-Instruments in der hier gewählten Lösung zu einer gemeinsamen Komponente, die die Qualität der Suchergebnisse beschreibt, zusammengefasst werden, lässt sich lediglich die vierte Skala, die sich auf die Benutzerfreundlichkeit der Suchmaschine bezieht, vollständig replizieren.

Einige mögliche Gründe für die gefundenen Abweichungen von der ursprünglichen Faktorstruktur sind bereits in Abschnitt 6.4.4.2 beschrieben. Neben der dort genannten Übersetzung des

Tab. 7.19.: Ergebnisse der Hauptkomponentenanalyse der EUCS-Items ($n = 128$). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Suchergebnis	Benutzerfreundlichkeit
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,96**	−0,09
F04	Liefert die Suchmaschine genügend Information?	**0,93**	−0,05
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,92**	0,00
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,90**	0,00
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,87**	0,01
F05	Ist die Suchmaschine präzise?	**0,83**	0,101
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	**0,69**	0,22
F09	Ist die Suchmaschine einfach zu bedienen?	−0,06	**0,94**
F08	Ist die Suchmaschine benutzerfreundlich?	0,13	**0,83**
	Skalenbezeichnung	SK12	SK13
	Anteil an Gesamtvarianz (in %)	60,96	19,41
	α	0,95	0,77

Tab. 7.20.: Ergebnisse der Hauptkomponentenanalyse der Zusatzitems ($n = 128$). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Suche	Eigenleistung	Aufgabe	Benutzerfreundlichkeit
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,89**	0,13	−0,04	−0,02
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	**0,78**	−0,21	0,09	0,26
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,59**	0,34	0,20	−0,08
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,10	**0,83**	−0,06	0,11
F23	Ich bin mit meiner Suchleistung zufrieden.	−0,04	**0,70**	0,25	0,14
F14	Es war einfach, die Aufgabe zu bearbeiten.	−0,02	−0,02	**0,98**	0,02
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	0,31	0,27	**0,51**	0,04
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	0,02	0,08	0,01	**0,93**
	Skalenbezeichnung	SK14	SK15	SK16	-
	Anteil an Gesamtvarianz (in %)	28,51	22,73	20,95	14,98
	α	0,86	0,79	0,84	-

Fragebogens und dem besonderen Kontext des Experiments kommt im Vergleich zum zweiten Experiment noch hinzu, dass diesmal die gemittelten Zufriedenheitswerte für jeden Teilnehmer in die Analyse einfließen. Dies führt einerseits dazu, dass die in die Analyse einbezogene Stichprobengröße im Fall des dritten Experiments deutlich geringer ausfällt (240 vs. 128), was eine mögliche Erklärung für die unterschiedlichen Ergebnisse der beiden Experimente darstellt, da laut Field et al. (2012, S. 769) Korrelationskoeffizienten besonders bei kleinen Stichprobenumfängen von Stichprobe zu Stichprobe variieren können. Andererseits werden Schwankungen in den Einzelmessungen durch die Mittelwertbildung reduziert und somit die Generalisierbarkeit der Ergebnisse erhöht.

Auch für die Analyse der acht nicht im EUCS-Instrument enthaltenen Items sind alle Kriterien zur Berechnung einer Faktorenanalyse erfüllt (vgl. Tab. 7.18). Für die Hauptkomponentenanalyse mit Oblimin-Rotation wird die Faktorenzahl des zweiten Experiments zugrunde gelegt. Obwohl zwei Items im dritten Experiment aus methodischen Gründen nicht mit erhoben werden (vgl. Abschn. 7.3.2), kann die ursprüngliche Faktorstruktur des zweiten Experiments überwiegend reproduziert werden. Das Ergebnis der Hauptkomponentenanalyse ist in Tabelle 7.20 aufgeführt, die Varianzaufklärung beträgt hier $87{,}1\,\%$. Im Vergleich zum zweiten Experiment gibt es bei zwei Subskalen Itemverschiebungen, die jedoch inhaltlich gerechtfertigt erscheinen. So bildet Item

Tab. 7.21.: Ergebnisse der Hauptkomponentenanalyse aller Items ($n = 128$). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Suche	Benutzer-freundlich-keit	Eigen-leistung-	Aufgabe
F04	Liefert die Suchmaschine genügend Information?	**0,91**	−0,07	0,03	0,03
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,87**	−0,13	0,06	0,11
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,82**	−0,11	0,16	0,05
F05	Ist die Suchmaschine präzise?	**0,75**	0,20	0,06	−0,03
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	**0,71**	0,29	−0,19	0,12
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,67**	0,38	0,04	−0,12
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,50**	0,14	0,27	0,15
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	0,09	**0,79**	0,00	0,14
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	0,00	**0,70**	0,28	0,05
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,11	0,05	**0,85**	−0,05
F23	Ich bin mit meiner Suchleistung zufrieden.	0,03	0,06	**0,70**	0,25
F14	Es war einfach, die Aufgabe zu bearbeiten.	0,04	0,04	0,02	**0,94**
	Skalenbezeichnung	SK17	SK18	SK19	-
	Anteil an Gesamtvarianz (in %)	39,32	15,75	15,58	11,08
	α	0,95	0,76	0,79	-

F16 nun gemeinsam mit Item F14 die neue Skala Aufgabe und Item F17 nimmt stattdessen den Platz von F16 in der Skala Suche ein. Vor allem letzteres verwundert nicht sonderlich, da F17 im zweiten Experiment bereits eine recht hohe Nebenladung auf dieser Komponente aufweist. Die internen Konsistenzen der Skalen sind ausnahmslos als gut zu bewerten (Cronbachs Alpha $> 0{,}79$).

In Bezug auf die Analyse aller Zufriedenheitsitems scheint wie schon im zweiten Experiment ein Multikollinearitätsproblem vorzuliegen (vgl. Tab. 7.18). In einem ersten Schritt soll trotz dieser Multikollinearität zunächst die Replizierbarkeit der im zweiten Experiment verwendeten Lösung untersucht werden. Unter Einbezug der dort enthaltenen Items zeigt sich bereits, dass die Validierung der Faktorenstruktur erfolgreich ist. So lassen sich durch die Hauptkomponentenanalyse erneut vier Faktoren abgrenzen, die inhaltlich weitgehend mit den im zweiten Experiment identifizierten Faktoren Suche, Benutzerfreundlichkeit, Aufgabe und Eigenleistung übereinstimmen. Das erwähnte Multikollinearitätsproblem lässt sich durch den zusätzlichen Ausschluss der Items F02 und F16 vermeiden. Eine inhaltlich gut interpretierbare Faktorenlösung (KMO-Kriterium: $0{,}95$; Bartlett-Test: $p < 0{,}001$) ergibt sich durch den weiteren Ausschluss von Item F08 (vgl. Anh. E.2.2). Inhaltlich ergeben sich auf diese Weise die gleichen Faktoren wie im zweiten Experiment, die zusammen $81{,}7\,\%$ der Gesamtvarianz erklären (vgl. Tab. 7.21). Die internen Konsistenzen erweisen sich mit Werten über $0{,}76$ ebenfalls als vergleichbar. Itemverschiebungen im Vergleich zum zweiten Experiment ergeben sich im Wesentlichen durch den Wegfall der Items F08 und F09, welche im dritten Experiment zusammen die Skala Benutzerfreundlichkeit bilden (vgl. Tab. 7.19). In der hier gezeigten Faktorenlösung wird dieser Faktor durch die Items F17 und F24 abgedeckt. Auch in diesem Fall ist der Skalenwechsel der Items F07 und F17 jedoch inhaltlich nachvollziehbar und für die Validierung der Faktorstruktur nicht problematisch. Zu beachten ist weiterhin, dass der Faktor Aufgabe nun lediglich das Item F14 umfasst, weswegen in diesem Fall keine eigenständige Skala eingeführt wird. Die Ergebnisse der Voranalysen sind in Anhang E.2.2 detailliert dargestellt. Über die bis hier beschriebenen Faktorskalen hinaus, werden, analog

zu Experiment 2, erneut die gemittelten Zufriedenheitsurteile getrennt nach EUCS- (SK-E-13) und Zusatzitems (SK-Z-13) sowie aller Zufriedenheitsitems zusammen (SK-G-13) gebildet. Des Weiteren werden für diejenigen Skalen aus Experiment 2, für die alle Frageitems vorhanden sind, die entsprechenden Mittelwerte ebenfalls in die Auswertung einbezogen. Eine Übersicht über alle in Experiment 3 verwendeten Skalen kann Tabelle D.1 in Anhang D.2 entnommen werden.

Zusammenfassend bleibt an dieser Stelle festzuhalten, dass die Daten des dritten Experiments nicht in dem Maße die theoretische Faktorenstruktur des EUCS-Instruments widerspiegeln, wie im zweiten Experiment. In Bezug auf die Einbeziehung der zusätzlich entwickelten Zufriedenheitsindikatoren ist die Zuordnung der Items zu den jeweiligen Faktoren jedoch über beide Experimente hinweg vergleichbar und erlaubt darüber hinaus eine inhaltlich klare Interpretation der Faktorenstruktur. Hinsichtlich der Gütekriterien zeigt sich, dass alle Skalen ausreichend reliabel und auch über die Zeit in der Lage sind, verschiedene Dimensionen der Benutzerzufriedenheit valide zu erfassen.

7.4.4.3. Reliabilitäts- und Validitätsanalyse

In diesem Abschnitt werden die gefundenen Skalen einer Reliabilitäts- und Validitätsanalyse unterzogen. Dabei werden Genauigkeit und Gültigkeit der einzelnen Skalen mit den in Abschnitt 6.4.4.3 bereits beschriebenen üblichen Verfahren ermittelt. Ähnlich wie beim zweiten Experiment liegt der Schwerpunkt wiederum auf einer Beurteilung der einzelnen Skalen im Hinblick auf die in Abschnitt 7.4.2 beschriebenen Teilstichproben. Die zugrunde liegende Annahme ist dabei erneut, dass stabile Gütekriterien als guter Indikator für die Eignung der Gesamtstichprobe SP_A angesehen werden können.

Die interne Konsistenz sowie die kriteriumsbezogene Validität der acht Skalen sind in Tabelle 7.22 getrennt nach Aufgabe und Datenqualitätsstufe aufgeführt. Die Gütekriterien der Skalen sind insgesamt zufriedenstellend. Nur für die Skalen SK13 und SK18 zeigen sich bei der ersten Aufgabe Werte für Cronbachs Alpha unterhalb der kritischen Grenze von 0,7. Die Kriteriumsvalidität der Skalen wird wie im zweiten Experiment mithilfe des Außenkriteriums bestätigt. Alle Korrelationen liegen hier über dem von Doll und Torkzadeh (1988, S. 264) vorgeschlagenen Schwellenwert von 0,4.

Darüber hinaus zeigen sich die berechneten Reliabilitäts- und Validitätskoeffizienten als sehr stabil im Vergleich zwischen der bereinigten (SP_B) und der weniger kontrollierten Stichprobe (SP_A). Die maximale Differenz tritt wiederum im Kontext der ersten Aufgabe auf. Für die Kriteriumsvalidität der beiden identischen Skalen SK15 und SK19 beträgt diese bspw. 0,08 und kann damit als sehr gering angesehen werden. Ähnlich geringe Abweichungen zwischen den ermittelten Reliabilitäts- und Validitätskoeffizienten ergeben sich beim Vergleich der drei Suchaufgaben. Die größte Differenz in Bezug auf die Kriteriumsvalidität besteht für die beiden identischen Skalen SK15 und SK19. In der Stichprobe SP_A beträgt diese 0,12.

Wie schon im zweiten Experiment wird in einem letzten Schritt überprüft, ob der Ausschluss bestimmter kritischer Fallgruppen einen Unterschied auf die Ergebnisse ausübt und inwiefern die hier diskutierten Gütekriterien hinsichtlich der Originalskalen des EUCS-Instruments erfüllt sind. Auch diese Ergebnisse unterstützen die Eignung der Gesamtstichprobe und stehen in aggregierter Form in Anhang E.2.3 zur Verfügung.

Zusammenfassend lässt sich daher festhalten, dass die in diesem Abschnitt diskutierten Güte-

Tab. 7.22.: Skalenreliabilität und Kriteriumsvalidität nach Datenqualität.

Skala	Beschreibung	#Items	Cronbachs Alpha		Kriteriumsvalidität	
			SP_A	SP_B	SP_A	SP_B
			$n = 128$	$n = 86$	$n = 128$	$n = 86$
			$A1$			
SK12	Suchergebnis	7	0,89	0,9	0,8	0,79
SK13	Benutzerfreundlichkeit	2	0,57	0,54	0,55	0,58
SK14	Suche	3	0,8	0,84	0,75	0,78
SK15/19	Eigenleistung	2	0,75	0,8	0,59	0,67
SK16	Aufgabe	2	0,78	0,76	0,66	0,67
SK17	Suche	7	0,89	0,89	0,81	0,82
SK18	Benutzerfreundlichkeit	2	0,67	0,69	0,61	0,67
			$A2$			
SK12	Suchergebnis	7	0,95	0,95	0,89	0,89
SK13	Benutzerfreundlichkeit	2	0,75	0,76	0,56	0,63
SK14	Suche	3	0,86	0,85	0,8	0,81
SK15/19	Eigenleistung	2	0,82	0,81	0,77	0,77
SK16	Aufgabe	2	0,89	0,88	0,77	0,79
SK17	Suche	7	0,95	0,94	0,89	0,9
SK18	Benutzerfreundlichkeit	2	0,74	0,75	0,73	0,73
			$A3$			
SK12	Suchergebnis	7	0,96	0,96	0,92	0,91
SK13	Benutzerfreundlichkeit	2	0,73	0,72	0,63	0,66
SK14	Suche	3	0,88	0,88	0,87	0,87
SK15/19	Eigenleistung	2	0,82	0,81	0,82	0,86
SK16	Aufgabe	2	0,88	0,88	0,84	0,85
SK17	Suche	7	0,95	0,95	0,91	0,9
SK18	Benutzerfreundlichkeit	2	0,77	0,81	0,82	0,84

kriterien die Reliabilität und Validität der ermittelten Skalen bestätigen. Darüber hinaus wird die Eignung der Gesamtstichprobe SP_A im dritten Experiment erneut durch die Tatsache gestützt, dass sich die berichteten Gütekriterien beim Vergleich verschiedener Aufgaben und Fallgruppen erneut als äußerst robust erweisen.

7.4.5. Auswertung der Benutzerzufriedenheit

Die im vorangegangenen Abschnitt dargestellte Faktorenanalyse zeigt, dass sich ein Großteil der aufgeklärten Varianz auf vier relativ robuste Zufriedenheitsfaktoren zurückführen lässt: Die Zufriedenheit mit der gestellten Aufgabe, die Funktionalität und Benutzerfreundlichkeit des Systems, die Zufriedenheit mit der eigenen Leistung (Selbstbewertung) sowie die Qualität des Sucherlebnisses. In diesem Abschnitt soll nun der Einfluss von Systemleistung und Erwartungshaltung auf die Zufriedenheitsreaktion der Testpersonen genauer untersucht werden. In die Betrachtung einbezogen werden dabei, wie im zweiten Experiment, sowohl die Einzelitems als auch die daraus gebildeten Zufriedenheitsskalen. Die Auswertung erfolgt analog zu dem in Abschnitt 7.4.3 beschriebenen Verfahren zunächst wieder mittels zweifaktorieller Varianzanalysen auf Basis der über alle drei Aufgaben gemittelten Zufriedenheitswerte. Zusätzlich findet in einem zweiten Analyseschritt erneut die Überprüfung dynamischer Effekte statt. Wie schon bei der Darstellung der Benutzerleistung werden im Folgenden die Befunde anhand ausgewählter Zufriedenheitsmaße erläutert. Weitere, die jeweilige Interpretation stützende, Items sind hingegen zur besseren Dokumentation der Ergebnisse in Klammern angegeben. Dabei gilt erneut, dass diese Auflistungen als Maß für die Stabilität der Effekte interpretiert werden sollten, die zum

Verständnis der Befunde jedoch nicht im einzelnen nachvollzogen werden müssen.

7.4.5.1. Varianzanalyse der Mittelwerte

Die Ergebnisse des ersten Analyseschritts sind in den Tabellen 7.23 und 7.24 zusammengefasst. Wie im Fall der Benutzerleistung sind in Tabelle 7.23 zunächst die signifikanten Unterschiede der Stichprobe $SP_{A,M}$ dargestellt. Tabelle 7.24 enthält die in der Stichprobe $SP_{B,M}$ zusätzlich auftretenden Effekte. Auch im Rahmen der Benutzerzufriedenheit werden ausschließlich Ergebnisse dargestellt, die in mindestens vier von fünf Stichproben einen signifikanten Effekt zeigen. Berichtet wird für jede Variable die Stichprobe mit dem signifikantesten Ergebnis, wobei in wenigen Ausnahmefällen zwei Ergebnisse dargestellt werden, wenn sich die Stichproben mit dem geringsten p-Wert bei zwei signifikanten Haupteffekten unterscheiden. Zur besseren Einordnung der Befunde ist über Fußnoten gekennzeichnet, welche in $SP_{A,M}$ auftretenden Effekte durch $SP_{B,M}$ bestätigt werden und welches Treatment jeweils zu einer höheren Zufriedenheit führt. Weiterführende Informationen sowie eine vollständige Dokumentation der Ergebnisse aus Stichprobe $SP_{B,M}$ sind in Anhang E.3 enthalten.

Einleitend werden im Folgenden zunächst einige allgemeine Beobachtungen beschrieben. So fällt bspw. auf, dass, anders als bei den Leistungsvariablen, für keines der Zufriedenheitsitems ein Einfluss beider unabhängiger Variablen eindeutig ausgeschlossen werden kann (vgl. Tab. E.14). Tatsächlich werden mit Ausnahme von F11 (*Liefert die Suchmaschine aktuelle Information?*) und SK15-F (Faktorwert Eigenleistung) alle Einzelitems sowie Zufriedenheitsskalen in beiden Stichproben signifikant. Auch bezüglich der Qualität des Einflusses fällt auf, dass die überwiegende Mehrheit der Effekte eindeutig signifikant ist. So gibt es keine Variable in der Stichprobe $SP_{B,M}$ und nur zwei Variablen in der Stichprobe $SP_{A,M}$, bei denen lediglich ein der Tendenz nach signifikantes Ergebnis vorliegt (vgl. F17 u. SK19-F Tab. E.23). Insgesamt betrachtet kann davon ausgegangen werden, dass die hier verwendeten Frageitems und Zufriedenheitsskalen ein sensibles Instrumentarium für die Überprüfung von System- und Erwartungseinfluss darstellen.

Tab. 7.23.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerzufriedenheit in $SP_{A,M}$. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen E.19 und E.23 in Anhang E.3 entnommen werden.

		System		Erwartung	
ID	**Beschreibung**	**S_G**	**S_S**	**E_H**	**E_N**
F01[a]	Liefert die Suchmaschine genau die Information, die Sie benötigen?	3,32	3,23	**3,49**[b]	**3,06**
F02[a]	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	3,31	3,3	**3,51**[b]	**3,1**
F03[a]	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	3,06	2,95	**3,25**[b]	**2,75**
F04[a]	Liefert die Suchmaschine genügend Information?	3,28	3,25	**3,56**[b]	**2,97**
F05[a]	Ist die Suchmaschine präzise?	3,05	2,85	**3,28**[b]	**2,62**
F06[a]	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	3,06	2,86	**3,25**[b]	**2,67**
F07[a]	Finden Sie die Präsentation der Ergebnisse hilfreich?	3,25	3,12	**3,43**[b]	**2,94**
F08[a]	Ist die Suchmaschine benutzerfreundlich?	3,93	3,76	**4,18**[b]	**3,51**
F09[a]	Ist die Suchmaschine einfach zu bedienen?	4,43	4,49	**4,64**[b]	**4,27**
F10[a]	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	3,69	3,69	**3,96**[b]	**3,42**

Fortsetzung auf nächster Seite

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.

[b] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.

[c] Entspricht auch den Skalen SK15-M und SK19-M.

[d] Entspricht auch der Skala SK18-M.

[e] Entspricht auch der Skala SK13-M.

Tab. 7.23 (Fortsetzung)

ID	Beschreibung	System		Erwartung	
		S_G	S_S	E_H	E_N
F12[a]	Ist die Suchmaschine erfolgreich?	3,36	3,39	**3,71**[b]	**3,05**
F13[a]	Sind Sie mit der Suchmaschine zufrieden?	3,34	3,25	**3,62**[b]	**2,97**
F14[a]	Es war einfach, die Aufgabe zu bearbeiten.	3,91	3,85	**4,05**[b]	**3,71**
F16[a]	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	3,41	3,4	**3,67**[b]	**3,14**
F17[a]	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	3,32	3,17	**3,48**[b]	**3,01**
F18[a]	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	3,39	3,11	**3,46**[b]	**3,04**
F19[a]	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**3,18**[b]	**2,89**	**3,29**[b]	**2,77**
F20[a]	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	2,73	2,73	**2,98**[b]	**2,48**
F22[a]	Ich bin mit den Suchergebnissen zufrieden.	3,35	3,23	**3,55**[b]	**3,03**
F23[a]	Ich bin mit meiner Suchleistung zufrieden.	3,4	3,32	**3,54**[b]	**3,18**
F24[a]	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	3,22	3,06	**3,4**[b]	**2,87**
F25[a]	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	3,24	3,16	**3,5**[b]	**2,89**
F26[a]	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	2,73	2,52	**3,01**[b]	**2,24**
SK01-M[a]	Genauigkeit (Mittelwert)	3,08	2,99	**3,35**[b]	**2,72**
SK02-M[a]	Inhalt (Mittelwert)	3,18	3,15	**3,44**[b]	**2,9**
SK03-M[a]	Benutzerfreundlichkeit (Mittelwert)	3,53	3,46	**3,8**[b]	**3,19**
SK04-M[a]	Suche (Mittelwert)	3,35	3,18	**3,5**[b]	**3,03**
SK07-M[ad]	Benutzerfreundlichkeit (Mittelwert)	3,24	3,11	**3,42**[b]	**2,93**
SK08-M[a]	Suche (Mittelwert)	3,3	3,14	**3,45**[b]	**2,99**
SK09-M[a]	Benutzerfreundlichkeit (Mittelwert)	3,49	3,32	**3,65**[b]	**3,16**
SK11-M[ac]	Eigenleistung (Mittelwert)	3,02	3,06	**3,23**[b]	**2,85**
SK12-M[a]	Suchergebnis (Mittelwert)	3,2	3,14	**3,4**[b]	**2,94**
SK12-F[a]	Suchergebnis (Faktorwert)	0,08	-0,05	**0,39**[b]	**-0,36**
SK13-F[a]	Benutzerfreundlichkeit (Faktorwert)	0,04	-0,1	**0,38**[b]	**-0,43**
SK14-M[a]	Suche (Mittelwert)	3,29	3,06	**3,39**[b]	**2,95**
SK14-F[a]	Suche (Faktorwert)	0,16	-0,13	**0,36**[b]	**-0,32**
SK16-M[a]	Aufgabe (Mittelwert)	3,68	3,57	**3,84**[b]	**3,4**
SK16-F[a]	Aufgabe (Faktorwert)	0,03	-0,13	**0,26**[b]	**-0,35**
SK17-M[a]	Suche (Mittelwert)	3,2	3,08	**3,41**[b]	**2,87**
SK17-F[a]	Suche (Faktorwert)	0,13	-0,07	**0,42**[b]	**-0,37**
SK18-F[a]	Benutzerfreundlichkeit (Faktorwert)	0,16	-0,17	**0,32**[b]	**-0,33**
SK19-F[a]	Eigenleistung (Faktorwert)	0	-0,07	**0,21**[b]	**-0,28**
SK-A[a]	Accuracy (EUCS)	3	2,91	**3,31**[b]	**2,61**
SK-C[a]	Content (EUCS)	3,26	3,19	**3,48**[b]	**2,97**
SK-E[ae]	Ease of Use (EUCS)	4,2	4,1	**4,39**[b]	**3,91**
SK-T[a]	Timeliness (EUCS)	3,58	3,55	**3,79**[b]	**3,34**
SK-K[a]	Kriteriumsskala	3,34	3,31	**3,67**[b]	**2,98**
SK-E-88[a]	EUCS-Skala-1988	3,5	3,41	**3,68**[b]	**3,22**
SK-E-09[a]	EUCS-Skala-2009	3,26	3,2	**3,5**[b]	**2,96**
SK-E-13[a]	EUCS-Skala-2013	3,38	3,3	**3,61**[b]	**3,07**
SK-Z-13[a]	Zusatzskala-2013	3,29	3,21	**3,47**[b]	**3,02**
SK-G-13[a]	Gesamtskala-2013	3,22	3,13	**3,44**[b]	**2,9**
E02[a]	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	3,41	3,22	**3,6**[b]	**3,03**
E03[a]	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	3,33	3,18	**3,54**[b]	**2,97**
E04[a]	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	3,14	3,05	**3,31**[b]	**2,88**
E05[a]	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	**3,17**[b] **3,13**[b]	**2,84** **2,86**	**3,32**[b] **3,31**[b]	**2,69** **2,68**

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.

[b] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.

[c] Entspricht auch den Skalen SK15-M und SK19-M.

[d] Entspricht auch der Skala SK18-M.

[e] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. 7.23 (Fortsetzung)

ID	Beschreibung	System S_G	System S_S	Erwartung E_H	Erwartung E_N
E06-M[a]	Erwartungsskala	3,28	3,13	**3,44**[b]	**2,97**

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.
[b] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.
[c] Entspricht auch den Skalen SK15-M und SK19-M.
[d] Entspricht auch der Skala SK18-M.
[e] Entspricht auch der Skala SK13-M.

Des Weiteren zeigt sich, dass die im dritten Experiment gewählte Methode zur Erwartungsmanipulation als erfolgreich angesehen werden kann. So lässt sich in $SP_{A,M}$ bspw. für alle Zufriedenheitsitems mit Ausnahme von F11 ein signifikanter Effekt der Erwartungshaltung nachweisen. Daneben zeigen auch die fünf Items, die die Erwartungshaltung der Probanden erfragen (E02 bis E06-M), eindeutig eine positive Abhängigkeit von der induzierten Erwartungshaltung. Demgegenüber fällt der Einfluss der Systemleistung allgemein geringer aus, wobei der Systemeinfluss in der besser kontrollierten Stichprobe $SP_{B,M}$ stärker zu Tage tritt. Interaktionseffekte zwischen Systemleistung und Erwartungshaltung sind hingegen weder eindeutig noch in der Tendenz zu beobachten.

Tab. 7.24.: In $SP_{B,M}$ neu hinzukommende signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerzufriedenheit. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Informationen zu Interaktionen und p-Werten können den Tabellen E.20 und E.24 in Anhang E.3 entnommen werden.

ID	Beschreibung	System S_G	System S_S	Erwartung E_H	Erwartung E_N
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	**3,43**[a]	**2,93**	**3,49**[a]	**2,87**
		3,43[a]	**3,04**	**3,53**[a]	**2,94**
SK07-M[b]	Benutzerfreundlichkeit (Mittelwert)	**3,33**[a]	**2,93**	**3,41**[a]	**2,85**
		3,33[a]	**3,01**	**3,47**[a]	**2,86**
SK14-M	Suche (Mittelwert)	**3,3**[a]	**2,93**	**3,43**[a]	**2,81**
SK14-F	Suche (Faktorwert)	**0,22**[a]	**-0,32**	**0,32**[a]	**-0,43**
		0,22[a]	**-0,3**	**0,36**[a]	**-0,44**
SK15-F	Eigenleistung (Faktorwert)	-0,14	0,01	**0,35**[a]	**-0,47**
SK18-F	Benutzerfreundlichkeit (Faktorwert)	**0,25**[a]	**-0,35**	**0,29**[a]	**-0,39**
		0,25[a]	**-0,33**	**0,33**[a]	**-0,41**
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	**3,44**[a]	**3,15**	**3,62**[a]	**2,98**

[a] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.
[b] Entspricht auch der Skala SK18-M.

Obwohl in diesem Teil der Auswertung durch die Mittelung über die drei Aufgaben keine zeitliche Information mehr explizit in den Daten vorhanden ist, bestätigen die Ergebnisse der Varianzanalysen inhaltlich die Befunde des zweiten Experiments. Sowohl eine höhere Erwartungshaltung der Probanden als auch eine bessere Systemqualität führen in allen Fällen zu einer höheren Zufriedenheit. In Bezug auf die Erwartung widersprechen die Ergebnisse somit wiederum der Vorhersage des C/D-Paradigmas, wonach in Bezug auf die Erwartungshaltung

das umgekehrte Verhalten zu erwarten wäre. Eine mögliche Erklärung ergibt sich aus dem in Abschnitt 7.4.3 dargestellten erwartungsgesteuerten Anpassungseffekt der Relevanzwahrnehmung. Da eine erhöhte Erwartungshaltung zu einer weniger restriktiven Relevanzbeurteilung führt, erscheint es nicht verwunderlich, dass in diesem Fall auch das gesamte System positiver wahrgenommen wird. Dies schlägt sich wiederum in einer höheren Zufriedenheit nieder. Die Zufriedenheit mit dem System könnte dadurch, wie schon bei Experiment 2 vermutet, von der Wahrnehmung der eigenen Suchleistung überlagert sein.

Hinsichtlich der Systemleistung werden die Annahmen des C/D-Paradigmas hingegen erneut bestätigt. So führt in Fällen, in denen sie signifikant wird, eine höhere Systemleistung zu einer größeren Zufriedenheit der Probanden. Der Systemunterschied wird also unabhängig von der durch die Erwartungshaltung bedingten Anpassung der Relevanzbewertung wahrgenommen. Obwohl insgesamt nur sieben unterschiedliche Frageitems und Skalen eine signifikante Abhängigkeit von der Systemleistung zeigen, decken diese jedoch viele der in Abschnitt 7.4.4.2 identifizierten Zufriedenheitsaspekte ab. Genauer zeigt sich ein positiver Einfluss der Systemleistung auf die Wahrnehmung der Benutzerfreundlichkeit (F17, SK07-M/SK18-M/F), der Precision der Suchmaschine (F19) sowie der Qualität des Sucherlebnisses (SK14-M/F), aber auch auf die nach jeder Aufgabe erfragte Erwartung der Testpersonen (E02 u. E05).

In Bezug auf die Stichproben $SP_{A,M}$ und $SP_{B,M}$ fällt auf, dass der Einfluss der Systemleistung im Vergleich zur manipulierten Erwartungshaltung weniger stabil erscheint. So können nur zwei der sieben signifikanten Systemhaupteffekte (F19 u. E05) schon für die weniger kontrollierte Stichprobe $SP_{A,M}$ nachgewiesen werden. Die übrigen Effekte hingegen werden erst im Kontext der Stichprobe $SP_{B,M}$ sichtbar. Neben dem Erwartungsitem E05 ist somit insbesondere die Zufriedenheit mit der Precision des Suchsystems signifikant. Dies steht im Einklang mit den Ergebnissen der Benutzerleistung, bei denen ebenfalls zu beobachten ist, dass vorrangig die Precisionmaße sensibel auf die Systemqualität reagieren. Ein weiterer Grund für den geringeren Einfluss der Systemqualität auf die Nutzerzufriedenheit könnte im systembedingten Anpassungseffekt der Relevanzwahrnehmung begründet liegen. Da Nutzer des schlechteren Suchsystems weniger strenge Relevanzkriterien anwenden, fällt die tatsächlich wahrgenommene Systemleistung höher aus, während im Fall der hohen Systemleistung der umgekehrte Effekt zu beobachten ist (vgl. Abb. 7.17). In diesem Sinne verringert die Adaption der Relevanzkriterien also den wahrgenommenen Systemunterschied, wodurch sich gleichzeitig das Zufriedenheitsempfinden der beiden Untersuchungsgruppen angleicht. Im Ergebnis ist somit nur ein geringer Einfluss der Systemqualität auf das Zufriedenheitsurteil zu beobachten.

Für die gemittelten Zufriedenheitsdaten kann somit zusammenfassend festgehalten werden, dass die Erwartungsmanipulation des dritten Experiments hervorragend funktioniert. Mit Ausnahme zweier Frageitems kann ein signifikanter Einfluss der Erwartungshaltung auf alle Zufriedenheitsfragen und -skalen nachgewiesen werden und es bestätigen sich im Wesentlichen die Resultate des zweiten Experiments. Ähnliches gilt für den Einfluss der Systemqualität, wenngleich sich hier die Effekte weniger stabil in Bezug auf die Stichproben $SP_{A,M}$ und $SP_{B,M}$ darstellen.

7.4.5.2. Dynamische Entwicklung der Benutzerzufriedenheit

In diesem Abschnitt werden die Ergebnisse in Bezug auf die dynamische Entwicklung der Nutzerzufriedenheit vorgestellt. Dabei geht nun, analog zur dynamischen Analyse der Benutzerleistung,

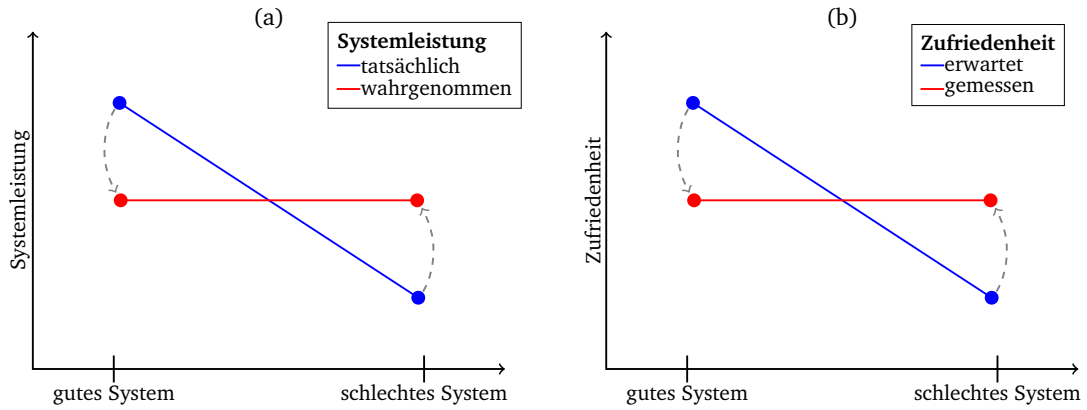


Abb. 7.17.: Einfluss des systembedingten Anpassungseffekts auf die Suchergebniswahrnehmung und die Benutzerzufriedenheit. Bild (a): Die gegenläufige Adaption der Relevanzwahrnehmung für die beiden Systemgüten führt zu einer Reduktion des wahrgenommenen Systemunterschieds im Vergleich zur tatsächlichen Systemgüte (gestrichelte Pfeile). Bild (b): Die Reduktion des wahrgenommenen Systemunterschieds impliziert eine Angleichung des Zufriedenheitsurteils.

die Aufgabenposition als zusätzlicher Faktor neben Systemleistung und Erwartungshaltung in die Auswertung mit ein. Eine ausführliche Beschreibung dieses Messwiederholungsdesigns kann dementsprechend Abschnitt 7.4.3.3 entnommen werden.

Es zeigt sich, dass die Benutzerzufriedenheit über alle drei Suchaufgaben hinweg als weitestgehend stabil anzusehen ist. Im Ergebnis ergibt sich lediglich für ein einzelnes Zufriedenheitsitem (F09: Ist die Suchmaschine einfach zu bedienen?) in der Stichprobe SP_A eine signifikante Abhängigkeit von der Aufgabenposition ($n = 72$, robuste Analyse, $F = 7{,}20$, $p = 0{,}0008$). Der Post-hoc-Test ergibt, dass sich die Zufriedenheit nach der ersten Suchaufgabe signifikant sowohl von der zweiten, als auch von der dritten Aufgabe unterscheidet ($P_1 = 4{,}7$, $P_2 = 4{,}3$, $P_3 = 4{,}3$). Die Zufriedenheit mit der Benutzerfreundlichkeit der Suchmaschine scheint also nach der ersten Aufgabe nachzulassen. Eine Darstellung der gesamten Teststatistik, sowie die übrigen Gruppenmittelwerte in Bezug auf Systemgüte und Erwartungshaltung können den Tabellen E.29 und E.42 in Anhang E.4 entnommen werden.

Im Folgenden werden nun die Effekte von Systemleistung und Erwartungshaltung unter Berücksichtigung der Aufgabenposition diskutiert. Die entsprechenden Mittelwerte und Teststatistiken sind aus Platzgründen wiederum in Anhang E.4 in die Tabellen E.29 und E.42 (SP_A) sowie E.30 und E.43 (SP_B) ausgelagert. Des Weiteren ist zu beachten, dass aufgrund von Topiceffekten lediglich für vier Zufriedenheitsindikatoren (F02, F12, SK11-M u. SK13-M/SK-E) eine ausreichende Anzahl an Fällen vorhanden ist, um eine Auswertung in der Stichprobe SP_B vornehmen zu können (vgl. Abschn. 7.4.2).

Auch unter Berücksichtigung der Aufgabenposition zeigt sich noch immer eine starke Abhängigkeit der Zufriedenheit von der Erwartungshaltung der Testteilnehmer. Hier können mit Ausnahme von sechs Zufriedenheitsindikatoren (F10, F17, F18, F20, SK07-M/SK18-M u. SK-T) in SP_A alle signifikanten Ergebnisse der Mittelwertanalyse bestätigt werden, für die die dynamische Auswertung durchgeführt werden kann. Nach wie vor führt eine höhere Erwartungshaltung zu einer größeren Zufriedenheit der Probanden. Darüber hinaus ergibt sich im dynamischen Fall für

keines der betrachteten Zufriedenheitsitems eine signifikante Systemabhängigkeit. Diese Instabilität der Systemeffekte kann als weiterer Hinweis darauf gedeutet werden, dass die Systemgüte im Zusammenspiel mit dem Anpassungseffekt der Relevanzwahrnehmung nur einen geringen Einfluss auf die Nutzerzufriedenheit ausübt (vgl. Abb. 7.18).

Im Gegensatz zu den Haupteffekten lassen sich im Kontext der Zufriedenheit keine tendenziell oder eindeutig signifikanten Interaktionen nachweisen. Um trotzdem einen Eindruck von der dynamischen Abhängigkeit der Nutzerzufriedenheit von Systemgüte und Erwartungshaltung zu gewinnen, wird an dieser Stelle analog zur dynamischen Auswertung der Benutzerleistung auf die Ergebnisse der klassischen Varianzanalyse zurückgegriffen. Da darüber hinaus auch im Kontext dieses Verfahrens mit größerer Teststärke keine über mindestens vier Stichproben hinweg stabilen Effekte nachweisbar sind, werden im Folgenden auch Zufriedenheitsvariablen in die Betrachtung mit einbezogen, für die sich nur in einzelnen Fällen signifikante Wechselwirkungseffekte zeigen. Zwar führt die geringere Stabilität dieser Effekte zu einer eingeschränkten Generalisierbarkeit der Resultate, jedoch lassen sich auf diese Weise Trends in Bezug auf die dynamische Entwicklung der Benutzerzufriedenheit identifizieren.

Interaktionseffekte zeigen sich in SP_A für die Wechselwirkung zwischen Erwartungshaltung und Systemleistung für die Variablen F03, SK-G-13, SK03-M, SK09-M und SK12-M, zwischen Systemleistung und Aufgabenposition für die Variablen F04, F08, F20, F23 und F26 sowie zwischen Erwartungshaltung und Aufgabenposition für die Variablen F04, F08, F13, F24, SK07-M, SK09-M, SK-T, SK-E, E03 und E06-M. Dreifachwechselwirkungen zwischen allen drei unabhängigen Variablen sind hingegen nicht zu beobachten. In SP_B lässt sich, vermutlich aufgrund des geringen Stichprobenumfangs, lediglich für SK-E eine Wechselwirkung zwischen Erwartungshaltung und Aufgabenposition nachweisen. Die zugehörigen Teststatistiken und Mittelwerte können den Tabellen E.36 bis E.39 in Anhang E.4.3 bzw. den Tabellen E.44 und E.45 in Anhang E.4.4 entnommen werden.

Thematisch decken die fünf Zufriedenheitsindikatoren, für die sich eine signifikante Wechselwirkung zwischen Systemleistung und Erwartungshaltung nachweisen lässt, die Dimensionen Benutzerfreundlichkeit (SK03-M u. SK09-M) sowie Qualität der Ergebnislisten (F03 u. SK12-M) ab. Des Weiteren ist der Effekt mit SK-G-13 in Bezug auf den Mittelwert aller im dritten Experiment untersuchten Zufriedenheitsitems sichtbar. Da das Verhalten in den vier Untersuchungsgruppen qualitativ für alle Variablen übereinstimmt, ist in Abbildung 7.18 das Interaktionsdiagramm beispielhaft für die allgemeine Zufriedenheitsskala SK-G-13 dargestellt. Zunächst lässt sich erkennen, dass einerseits bei hoher Erwartungshaltung die Systemgüte keinen weiteren Einfluss auf die Zufriedenheit der Benutzer auszuüben scheint (Bild (b)) und umgekehrt im Fall des besseren Systems die Erwartungshaltung der Probanden unerheblich für das abschließende Zufriedenheitsurteil ist (Bild (a)). Demgegenüber besteht ein deutlicher Unterschied zwischen gutem und schlechtem System in der Untersuchungsgruppe mit niedriger Erwartungshaltung bzw. in Bezug auf hohe und niedrige Erwartungshaltung im Fall des schlechten Systems. Hinsichtlich des erwartungsgesteuerten Anpassungseffekts lassen sich diese Befunde dergestalt interpretieren, dass dieser im Wesentlichen nur im Fall der geringeren Systemleistung zu beobachten ist, wobei wiederum eine hohe Erwartungshaltung zu einer größeren Zufriedenheit und zwar in etwa auf dem Niveau der besseren Systemleistung führt. Dieses Verhalten widerspricht jedoch erneut

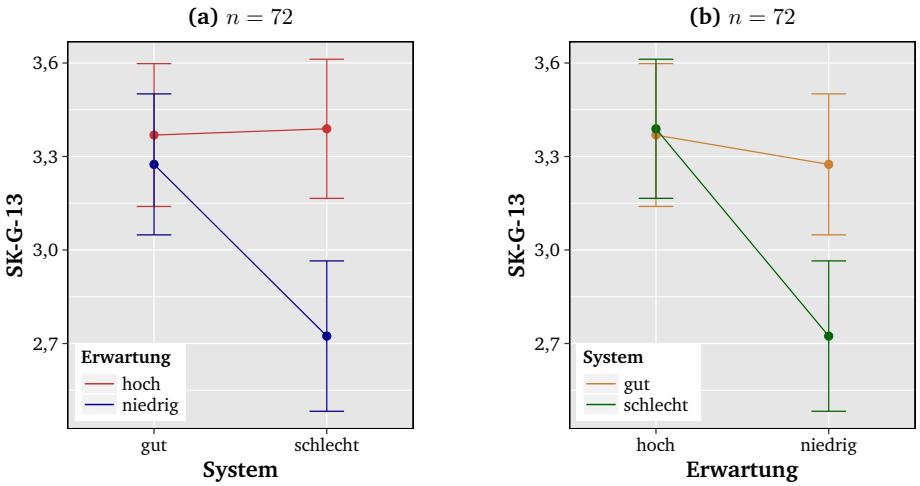


Abb. 7.18.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für die gemittelte Gesamtzufriedenheitsskala (SK-G-13). Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Unterschiede in Bezug auf die Erwartungshaltung lassen sich im Wesentlichen bei Probanden des schlechteren Systems beobachten (a), während ein Unterschied in Bezug auf die Systemleistung ausschließlich im Kontext der niedrigen Erwartungshaltung auftritt (b). Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte.

den Vorhersagen des C/D-Paradigmas, wonach enttäuschte Erwartungen zu einer geringeren Zufriedenheit führen sollten. Demgegenüber ist für die bessere Systemleistung der Einfluss der Erwartungshaltung vernachlässigbar. Der Unterschied zwischen den beiden Systemleistungen wiederum wird von den Probanden nur im Fall der niedrigen Erwartungshaltung wahrgenommen und stimmt hier mit den Vorhersagen des C/D-Paradigmas überein, wonach eine Übererfüllung der Erwartung zu einer größeren Zufriedenheit führen sollte. Die Abwesenheit des Systemeinflusses im Fall der hohen Erwartungshaltung lässt sich hingegen als Folge des systembedingten Anpassungseffekts der Relevanzwahrnehmung interpretieren: Wie in Abbildung 7.12 in Abschnitt 7.4.3.2 beschrieben, ist eine strengere, die wahrgenommenen Systemunterschiede verringernde Relevanzbeurteilung hauptsächlich im Fall der hohen Erwartungshaltung zu beobachten. In diesem Fall nehmen die Teilnehmer die beiden Systeme also als gleichwertig wahr und in Folge dessen zeigen sie auch eine vergleichbare Zufriedenheitsreaktion.

Im Folgenden wird genauer auf die Wechselwirkungen der Aufgabenposition mit der Systemgüte bzw. der Erwartungshaltung eingegangen. Dazu sind in Abbildung 7.19 stellvertretend die entsprechenden Interaktionsgraphen für das Frageitem F04 (*Liefert die Suchmaschine genügend Information?*) dargestellt, für das beide Wechselwirkungen signifikant werden. In Bezug auf die Systemleistung zeigen die Zufriedenheitsindikatoren F08 und F26, die Erwartungshaltung betreffend die Zufriedenheitsmaße F08, F24, SK09-M, SK-E und SK-T sowie die beiden Erwartungsindikatoren E03 und E06-M ein qualitativ ähnliches Verhalten. Mit Blick auf die thematische Ausrichtung dieser Variablen zeigt sich, dass beide Wechselwirkungen die Themenbereiche Suchleistung und Benutzerfreundlichkeit umfassen. In allen Fällen ist zu beobachten, dass die Zufriedenheitsunterschiede im Hinblick auf die Systemgüte bzw. die Erwartungshaltung mit der

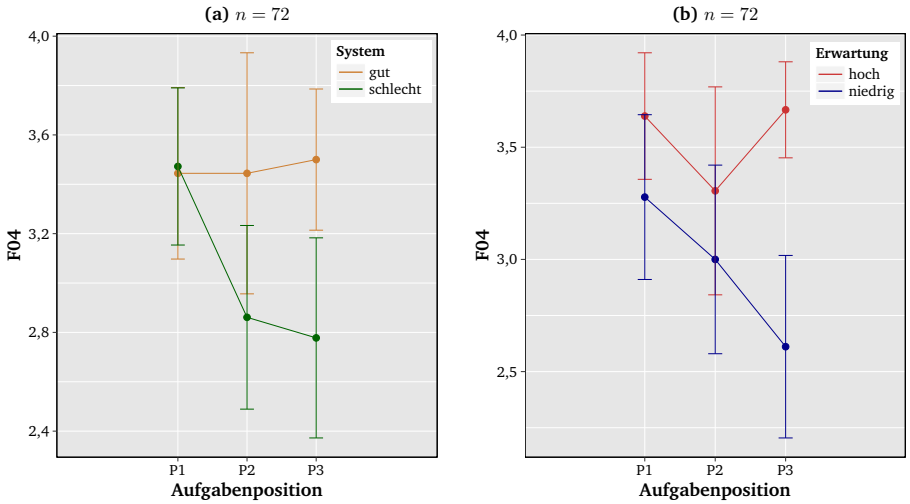


Abb. 7.19.: Wechselwirkung zwischen Systemleistung (a) bzw. Erwartungshaltung (b) und Aufgabenposition für das Zufriedenheitsitem: *Liefert die Suchmaschine genügend Information?* (F04). In beiden Fällen ist eine Zunahme der Untersuchungsgruppenunterschiede in Abhängigkeit von der Aufgabenposition zu erkennen. Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte.

Aufgabenposition zunehmen (vgl. Bild (a) bzw. Bild (b) in Abb. 7.19). Im Widerspruch zu Forschungshypothese H4 klingt also insbesondere der Erwartungseinfluss nicht ab sondern verstärkt sich vielmehr noch über die Suchaufgaben hinweg. Ein mögliches Erklärungsmodell für dieses Verhalten liefert erneut der erwartungsinduzierte Anpassungseffekt. So führt eine anfänglich negative Erwartungshaltung bei der ersten Aufgabe zu einer negativeren Relevanzwahrnehmung und somit ebenfalls zu einer geringeren Zufriedenheit. Diese impliziert darüber hinaus eine negative Erwartungshaltung für die nächste Aufgabe, was erneut eine geringere Zufriedenheit nach sich zieht. Betrachtet man die paarweisen Mittelwertunterschiede zwischen den sechs Versuchsgruppen, fällt auf, dass sich der Unterschied in Bezug auf die Erwartungshaltung bei der dritten Aufgabenposition noch einmal zu verstärken scheint. Insbesondere zeigt sich dieses Verhalten in gleicher Weise für die Erwartungsitems E03 und E06-M. Auch hier ist der Unterschied in Abhängigkeit von der initialen Erwartungshaltung bei der dritten Aufgabe am größten, was diese Interpretation noch einmal untermauert.

Im Gegensatz zur Erwartungshaltung scheinen die Ergebnisse in Bezug auf die Systemleistung dem Anpassungseffekt auf den ersten Blick zu widersprechen. So führt die strengere Relevanzbewertung im Kontext der höheren Systemgüte zunächst nicht zu einer geringeren Zufriedenheit der Probanden. Vielmehr nimmt die Zufriedenheit mit dem schlechteren System, das eigentlich als besser wahrgenommen werden müsste, über die Zeit ab. Um dieses Verhalten besser zu verstehen ist in Abbildung 7.20 zusätzlich ein Interaktionsdiagramm für alle zwölf Untersuchungsgruppen angegeben, wenngleich die Dreifachwechselwirkung zwischen Aufgabenposition, Systemgüte und Erwartungshaltung nicht signifikant ist. Jedoch ist in den beiden Interaktionsgraphen gut zu erkennen, dass die Abnahme der Zufriedenheit in Bezug auf die Systemgüte allein bei Nutzern mit der niedrigen Erwartungshaltung auftritt und sich bei diesen von Aufgabe zu Aufgabe verringert. Bei Probanden mit der hohen Erwartungshaltung zeigt sich hingegen im Wesentlichen

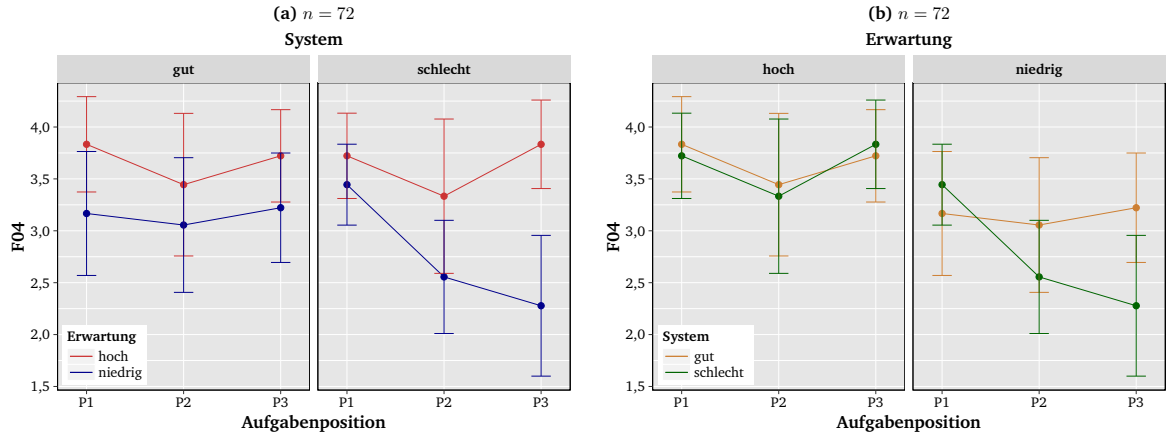


Abb. 7.20.: Nicht signifikante Wechselwirkung zwischen Systemleistung, Erwartungshaltung und Aufgabenposition für das Zufriedenheitsitem: *Liefert die Suchmaschine genügend Information?* (F04). Während im Kontext der hohen Erwartungshaltung (a) und der besseren Systemqualität (b) jeweils nur eine geringe dynamische Abhängigkeit zu beobachten ist, nehmen die Gruppenunterschiede in den anderen beiden Fällen (niedrige Erwartungshaltung bzw. schlechtes System) für die jeweils andere unabhängige Variable über die Zeit zu. In beiden Fällen ist eine Zunahme der Untersuchungsgruppenunterschiede in Abhängigkeit von der Aufgabenposition zu erkennen. Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte.

keine dynamische Änderung der Zufriedenheit in Bezug auf die Systemgüte. Dieses Verhalten erklärt sich wiederum aus dem systembedingen Anpassungseffekt, der, wie in Abbildung 7.12 in Abschnitt 7.4.3.2 zu erkennen ausschließlich im Kontext einer hohen Erwartungshaltung auftritt und somit eine Wahrnehmung des Systemunterschieds durch Probanden mit einer hohen Erwartungshaltung unterbindet. Interessant ist in diesem Zusammenhang darüber hinaus, dass die Dynamik in Bezug auf die Zufriedenheit erneut ausschließlich beim schlechteren System zu verorten ist. Während die Zufriedenheit mit dem besseren System weitestgehend konstant bleibt, nimmt die Zufriedenheit mit dem schlechteren System kontinuierlich ab.

Die verbleibenden beiden Zufriedenheitsitems (F13 u. SK07-M), die eine signifikante Wechselwirkung zwischen Erwartungshaltung und Aufgabenposition erkennen lassen, zeigen zwar ein zu F04 analoges Verhalten, hier fällt der Unterschied in Bezug auf die Erwartungshaltung für die zweite beobachtete Aufgabe allerdings geringer aus, bevor er bei der dritten Aufgabe wiederum ansteigt (vgl. Abb. 7.21 Bild (a)). Die Zufriedenheitsitems F20 und F23 mit einer Wechselwirkung zwischen Systemgüte und Aufgabenposition zeigen im Vergleich zu F04 hingegen ein auch qualitativ anderes Verhalten. Hier nimmt in beiden Fällen der von der Systemqualität abhängige Zufriedenheitsunterschied über die Zeit ab und das Zufriedenheitsurteil fällt bei der ersten Aufgabe für das schlechtere System besser aus (vgl. Abb. 7.21 Bild (b)). Interessanterweise stehen beide Frageitems in Zusammenhang mit der Zufriedenheit in Bezug auf die eigene Suchleistung (F20: *Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.*, F23: *Ich bin mit meiner Suchleistung zufrieden.*). Hier könnte gerade das geringere Angebot relevanter Dokumente in den Ergebnislisten dazu führen, dass die Nutzer des schlechteren Systems ihre eigene Suchleistung in Bezug auf den Recall (F20) positiver einschätzen, als Probanden, denen das bessere Suchsystem präsentiert wird. Mit der zweiten Suchaufgabe hingegen scheint

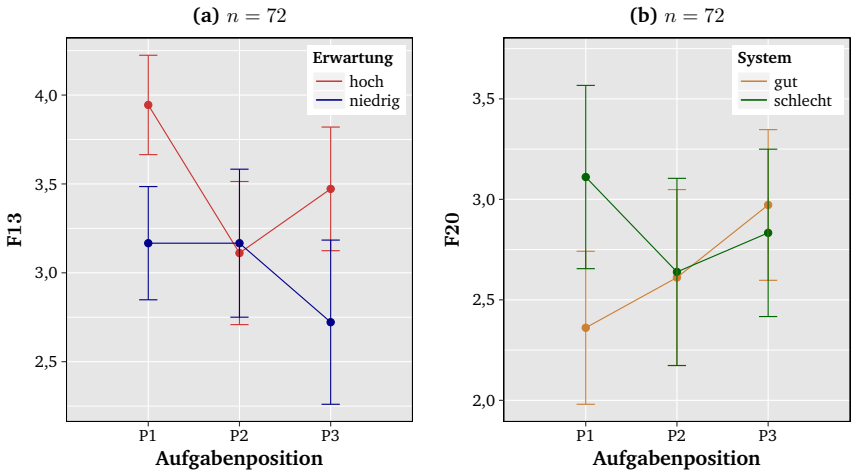


Abb. 7.21.: Von F04 abweichendes dynamisches Verhalten für die Zufriedenheitsitems F13 und F20. Bild (a) zeigt den Interaktionsgraph der signifikanten Wechselwirkung zwischen Erwartungshaltung und Aufgabenposition für das Zufriedenheitsitem: *Sind Sie mit der Suchmaschine zufrieden?* (F13). Hier fällt der Unterschied zwischen den Gruppenmittelwerten bei Aufgabe 2 besonders gering aus. Bild (b) stellt den Interaktionsgraph der signifikanten Wechselwirkung zwischen Systemleistung und Aufgabenposition für das Zufriedenheitsitem: *Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.* (F20) dar. Hier ist ein Nachlassen der Gruppenunterschiede über die Zeit zu beobachten, wobei initial die Zufriedenheit mit dem schlechteren System größer auszufallen scheint. Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte.

sich diese Wahrnehmung jedoch zu relativieren und es ist kein Unterschied mehr zwischen den beiden Systemgüten zu erkennen. Vor diesem Hintergrund dieser potentiell abweichenden Dynamik wird deutlich, dass die Erhebung der Wahrnehmung der eigenen Suchleistung im Kontext von IR-Experimenten nicht vernachlässigt werden sollte.

Zusammenfassend lässt sich festhalten, dass die Abhängigkeit der Zufriedenheit von der Erwartungshaltung der Probanden auch im Rahmen der dynamischen Auswertung bestätigt werden kann. Der Einfluss der Systemleistung stellt sich jedoch als weniger stabil heraus. Obwohl keine stabil signifikanten Wechselwirkungen zwischen den Haupteffekten vorliegen, zeigt sich, dass darüber hinaus die dynamische Entwicklung der Zufriedenheit eng mit den Anpassungseffekten verbunden zu sein scheint, wenngleich diese Ergebnisse aufgrund der fehlenden Stabilität nicht direkt generalisierbar sind.

7.4.6. Überprüfung der Gütekriterien des Experiments

Analog zu dem methodischen Vorgehen der beiden ersten Nutzerstudien wird im Folgenden die Stabilität der Ergebnisse der Hauptauswertung des dritten Experiments im Hinblick auf eine Reihe möglicher Störeinflüsse untersucht. Dabei lassen sich die betrachteten Störfaktoren erneut in zwei Gruppen untergliedern, abhängig davon, ob sie auf das untersuchungsmethodische Vorgehen zurückzuführen sind oder primär mit den individuellen Merkmalen der Probanden in Zusammenhang stehen. Insgesamt liefern die folgenden Analysen die Basis für eine angemessene Bewertung der im Hauptteil dieses Kapitels erhaltenen Forschungsergebnisse und lassen sich somit als Gütekriterien des Experiments verstehen.

7.4.6.1. Untersuchungsbedingte Störfaktoren

Den Schwerpunkt dieses Abschnitts bildet die Analyse möglicher Aufgabeneffekte sowie des Einflusses problematischer Fallgruppen, die im Rahmen der in Abschnitt 7.4.1 beschriebenen Stichprobenanalyse identifiziert werden. Konkret handelt es sich dabei um Testpersonen, die entweder unspezifisch suchen (SP_{SB}), möglicherweise das Testdesign durchschaut haben (SP_{TD}), die maximale Bearbeitungszeit unterschreiten (SP_{UZ}) oder bei denen die Erwartungsmanipulation versagt haben könnte (SP_{MV}). Da zu vermuten ist, dass sich der Einfluss dieser Störfaktoren schon auf der Ebene der einzelnen Testaufgaben manifestiert, erscheint es sinnvoll sie im Kontext der Aufgabeneffekte zu betrachten. Analog zur Überprüfung untersuchungsbedingter Störfaktoren im Rahmen des zweiten Experiments (vgl. Abschn. 6.4.6.1) wird dazu überprüft, ob der Ausschluss oben genannter Fallgruppen das Vorhandensein von Topiceffekten grundlegend verändert. Da für den letzten Störfaktor (SP_{UZ}) nur 33 der 153 Probanden die volle Bearbeitungsdauer ausschöpfen, wird in diesem Fall die Gruppe der Testpersonen, die bei jeder Aufgabe die volle Bearbeitungszeit von zehn Minuten ausschöpfen (SP_{IZ}), von der Analyse ausgeschlossen, um jeweils den größtmöglichen Stichprobenumfang zur Verfügung zu haben.

Um den in Abschnitt 7.4.1 vermuteten Einfluss der unterschiedlichen Testaufgaben zu untersuchen, werden zunächst, wie in Abschnitt 7.4.2 erläutert, einfaktorielle Varianzanalysen mit dem Suchthema als Messwiederholungsfaktor durchgeführt (vgl. Abschn. 4.3.2.4). Die unabhängigen Variablen hingegen bilden die im Rahmen der Hauptanalyse untersuchten Leistungsmaße und Zufriedenheitsindikatoren. Im Hinblick auf eine möglichst allgemein gültige Betrachtung des Untersuchungsgegenstandes wird bei der Fallauswahl darauf geachtet, dass pro Versuchsgruppe jedes Topic an allen drei Messpositionen gleich häufig vorkommt. Zur Umsetzung dieser Balancierung wird ein analoges Vorgehen wie bei der Topicbalancierung im Kontext von Experiment 2 verwendet (vgl. Abb. 6.16) und für jede unabhängige Variable wiederum fünf Zufallsstichproben gezogen. Um alle möglichen Effekte zu identifizieren und so eine möglichst konservative Überprüfung des Aufgabeneffekts zu gewährleisten, wird wie im zweiten Experiment auf die Vorgabe einer Mindeststichprobengröße verzichtet. Abhängig vom Ausgang der Voraussetzungsprüfung wird entweder die klassische oder eine robusten Variante der ANOVA mit Messwiederholung durchgeführt (vgl. Abschn. 4.3.2.4). Im Folgenden werden ausschließlich signifikante Ergebnisse, die in mindestens vier Zufallsstichproben nachweisbar sind, dargestellt. Da für die Stichproben mit signifikantem Topiceffekten nur in den seltensten Fällen die statistischen Voraussetzungen für die Anwendung eines klassischen varianzanalytischen Verfahrens erfüllt sind, werden im Folgenden aus Gründen der Übersichtlichkeit immer die Ergebnisse der robusten Post-Hoc-Tests berichtet. Des Weiteren dient dieser Analyseschritt der Identifikation derjenigen unabhängigen Variablen, die eindeutig, d.h. in allen fünf Zufallsstichproben, keinen Topiceffekt aufweisen und deshalb im Rahmen der Hauptauswertung keiner Topicbalancierung bedürfen (SP_{OT}). Eine Zusammenfassung der entsprechenden Leistungsmaße und Zufriedenheitsindikatoren findet sich in Tabelle E.48 in Anhang E.5.

Signifikante Unterschiede zeigen sich erneut hauptsächlich in der Gesamtstichprobe SP_A. Hier weisen vierzehn Leistungs- und sechzehn Zufriedenheitsvariablen einen signifikanten Aufgabeneffekt auf. Demgegenüber lassen sich in der bereinigten Stichprobe SP_B nur drei signifikante Aufgabeneffekte für die Benutzerleistung nachweisen, von denen zwei die Ergebnisse aus SP_A

bestätigen. Die Zufriedenheitsindikatoren hingegen zeigen für SP_B keine signifikante Topicabhängigkeit. Dies könnte als Hinweis gewertet werden, dass es sich bei SP_B tatsächlich um eine gut kontrollierte Stichprobe handelt. Da die in SP_B zur Verfügung stehenden Stichprobenumfänge jedoch lediglich 24 Probanden betragen, kann allerdings nicht ausgeschlossen werden, dass die Abwesenheit signifikanter Topiceffekte zu einem Teil auch auf die geringere Stichprobengröße zurückzuführen ist.

Um den relativen Schwierigkeitsgrad der Aufgaben untereinander zu ermitteln, werden im Anschluss an die Varianzanalyse Post-Hoc-Tests zur Bestimmung signifikanter Mittelwertsunterschiede zwischen den drei Suchaufgaben durchgeführt. Ihre Ergebnisse sind in den Tabellen 7.25 und 7.26 dargestellt. Da es sich um ein Messwiederholungsdesign handelt, entsprechen die angegebenen Stichprobenumfänge der pro Suchthema vorhandenen Probandenzahl. Zum besseren Verständnis der Tabellen bedarf es einiger Erläuterungen: Aus Übersichtsgründen wird pro Variable nur der Werteverlauf der Stichprobe mit dem signifikantesten Ergebnis exemplarisch dargestellt. Um dennoch einen Eindruck von der Qualität des jeweiligen Topiceffekts zu erhalten, wird per Fußnote kenntlich gemacht, welche Unterschiede nicht über alle fünf Stichproben hinweg Bestand haben. Die Abkürzung *d* bezeichnet die Differenz zwischen den Mittelwerten der verglichenen Suchthemen. Eine negative Differenz bedeutet, dass der Mittelwert des ersten Themas niedriger als der des zweiten Themas ist. Eine positive Differenz hingegen zeigt an, dass der erste Mittelwert höher ausfällt. In den Spalten *lwr* und *upr* sind die Unter- und Obergrenze der $95\,\%$ Konfidenzintervalle aufgeführt. Ein Effekt gilt als signifikant, wenn das Konfidenzintervall den Wert 0 nicht beinhaltet. Um darüber hinaus die Lesbarkeit der Ergebnisse zu erleichtern, kennzeichnet eine weitere Fußnote das jeweils leichtere (Tab. 7.25) bzw. zu einer höheren Benutzerzufriedenheit führende Thema (Tab. 7.26).

Im Folgenden werden zunächst die Befunde für die Benutzerleistung und anschließend die Resultate für die Benutzerzufriedenheit dargestellt. Die Ergebnisse aus Tabelle 7.25 zeigen, dass durchaus Unterschiede zwischen den drei Aufgaben vorhanden sind. So ergeben die Post-hoc-Tests für zehn der vierzehn signifikanten Leistungsmaße einen mindestens in der Tendenz vorhandenen signifikanten Unterschied im Schwierigkeitsgrad zwischen zwei Topics. In drei Fällen gilt dies sogar für die Unterschiede zwischen zwei der drei Topickombinationen (M35, B10 u. V51). In vier Fällen hingegen ist keine der Paarungen wenigstens in der Tendenz signifikant (M36,V09, V52 u. V54).

Es zeigt sich, dass der Schwierigkeitsgrad der einzelnen Aufgaben in Bezug auf unterschiedliche Leistungsmaße variieren kann. So stellt sich bspw. die Englischaufabe bei zwei Leistungsmaßen, die beide im Zusammenhang mit der Anzahl falsch als relevant bewerteter Dokumente stehen (M35 u.V51), als das leichteste Topic, d. h. signifikant leichter als die Wind- und die Wikiaufgabe, heraus. Darüber hinaus ist die Englischaufgabe in drei weiteren Fällen signifikant leichter als eines der beiden anderen Topics. Für andere Leistungsmaße hingegen stellt sich das Windtopic (B10) oder das Wikitopic (V34, V35 u. B10) als leichter als die Englischaufgabe heraus. Des Weiteren scheint es den Probanden für die Wikiaufgabe leichter gefallen zu sein Dokumente in Übereinstimmung mit den Juroren zu bewerten, als bei der Windaufgabe (M14). Insgesamt gesehen zeigt sich somit insbesondere, dass der Schwierigkeitsgrad einer Aufgabe kein absoluter Begriff ist, sondern verschiedene Aspekte einer Aufgabe, wie bspw. die Identifikation relevanter

Tab. 7.25.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung in SP_A. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die über alle fünf Stichproben hinweg nachweisbar sind.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
M14	Anz. richtig bew. Dok.	72	$< 0{,}05$[b]	Wind–Englisch[a]	-1,73	-3,44	-0,01
			$\mathbf{< 0{,}05}$	Wind–Wiki[a]	-2,16	-3,80	-0,52
			$> 0{,}05$	Englisch–Wiki	-0,43	-2,05	1,19
M35	Anz. falsch rel. bew. Dok. (4-st.)	72	$\mathbf{< 0{,}05}$	Wind–Englisch[a]	0,61	0,26	0,97
			$> 0{,}05$	Wind–Wiki	0,07	-0,39	0,53
			$\mathbf{< 0{,}05}$	Englisch[a]–Wiki	-0,55	-0,97	-0,13
M36	Anz. irrel. bew. Dok. (4-st.)	72	$< 0{,}05$[b]	Wind[a]–Englisch	-0,89	-1,50	-0,27
			$< 0{,}05$[b]	Wind[a]–Wiki	-0,57	-1,09	-0,04
			$> 0{,}05$	Englisch–Wiki	0,32	-0,42	1,05
B10	Durchschn. Bew. eher rel. Dok. (4-st.)	48	$\mathbf{< 0{,}05}$	Wind[a]–Englisch	1,27	0,48	2,06
			$> 0{,}05$	Wind–Wiki	0,49	-0,26	1,24
			$> 0{,}05$[c]	Englisch–Wiki[a]	-0,78	-1,61	0,05
V09	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	72	$< 0{,}05$[b]	Wind–Englisch[a]	0,08	0,02	0,13
			$> 0{,}05$	Wind–Wiki	0,04	-0,02	0,10
			$> 0{,}05$	Englisch–Wiki	-0,03	-0,07	0,001
V34	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$ (4-st.)	72	$> 0{,}05$[d]	Wind–Englisch	0,01	-0,003	0,03
			$> 0{,}05$	Wind–Wiki	-0,01	-0,03	0,001
			$\mathbf{< 0{,}05}$	Englisch–Wiki[a]	-0,03	-0,04	-0,01
V35	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$ (4-st.)	72	$< 0{,}05$[b]	Wind[a]–Englisch	0,03	0,005	0,05
			$> 0{,}05$	Wind–Wiki	-0,01	-0,04	0,02
			$\mathbf{< 0{,}05}$	Englisch–Wiki[a]	-0,04	-0,07	-0,01
V37	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	$\mathbf{< 0{,}05}$	Wind–Englisch[a]	-0,11	-0,20	-0,03
			$> 0{,}05$	Wind–Wiki	-0,02	-0,09	0,05
			$< 0{,}05$[b]	Englisch[a]–Wiki	0,09	0,01	0,17
V38	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$(4-st.)	72	$\mathbf{< 0{,}05}$	Wind–Englisch[a]	-0,02	-0,04	-0,01
			$> 0{,}05$	Wind–Wiki	-0,01	-0,02	0,01
			$> 0{,}05$	Englisch–Wiki	0,01	-0,001	0,03
V51	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$(4-st.)	72	$\mathbf{< 0{,}05}$	Wind–Englisch[a]	0,08	0,04	0,13
			$> 0{,}05$	Wind–Wiki	0,02	-0,03	0,07
			$\mathbf{< 0{,}05}$	Englisch[a]–Wiki	-0,06	-0,11	-0,02
V52	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$(4-st.)	24	$< 0{,}05$[b]	Wind–Englisch[a]	0,26	0,03	0,50
			$> 0{,}05$	Wind–Wiki	-0,10	-0,37	0,16
			$< 0{,}05$[b]	Englisch[a]–Wiki	-0,37	-0,54	-0,19
V54	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$(4-st.)	72	$> 0{,}05$[d]	Wind–Englisch	-0,08	-0,16	0,004
			$< 0{,}05$[b]	Wind[a]–Wiki	-0,10	-0,17	-0,02
			$> 0{,}05$	Englisch–Wiki	-0,02	-0,10	0,06
V79	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$(4-st.)	24	$> 0{,}05$	Wind–Englisch	-0,03	-0,07	0,003
			$> 0{,}05$	Wind–Wiki	0,01	-0,03	0,04
			$\mathbf{< 0{,}05}$	Englisch[a]–Wiki	0,04	0,01	0,06
S01	Anz. Suchen	72	$\mathbf{< 0{,}05}$	Wind–Englisch	-0,64	-1,05	-0,22
			$> 0{,}05$	Wind–Wiki	-0,20	-0,62	0,21
			$> 0{,}05$	Englisch–Wiki	0,43	-0,08	0,95

[a] Dieses Thema entspricht der leichteren Aufgabe.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.

oder irrelevanter Dokumente unterschiedlich schwer ausfallen können. Im Kontext der 4-stufigen Relevanzskala zeigt sich so bspw. dass es den Probanden für das Englischthema leichter fällt relevante Dokumente zu identifizieren (z.B. V37) und fälschlicherweise als relevant bewertete Dokumente zu vermeiden (z.B. V51). Beim Wikithema hingegen scheinen die Testteilnehmer einen größeren Anteil der zurückgegebenen eher relevanten Dokumente aufzurufen (V35).

Tab. 7.26.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit in SP_A. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	72	**< 0,05**	Wind[a]–Englisch	0,73	0,24	1,21
			> 0,05	Wind–Wiki	0,45	-0,02	0,93
			> 0,05	Englisch–Wiki	-0,27	-0,71	0,17
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	72	**< 0,05**	Wind[a]–Englisch	0,61	0,23	1,00
			> 0,05	Wind–Wiki	0,27	-0,10	0,64
			> 0,05	Englisch–Wiki	-0,34	-0,79	0,11
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	72	**< 0,05**	Wind[a]–Englisch	0,80	0,33	1,26
			> 0,05	Wind–Wiki	0,05	-0,33	0,42
			< 0,05[b]	Englisch–Wiki[a]	-0,75	-1,17	-0,33
F22	Ich bin mit den Suchergebnissen zufrieden.	72	**< 0,05**	Wind[a]–Englisch	0,77	0,38	1,16
			> 0,05	Wind–Wiki	0,07	-0,32	0,46
			< 0,05[b]	Englisch–Wiki[a]	-0,70	-1,25	-0,16
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	72	< 0,05[b]	Wind[a]–Englisch	0,55	0,07	1,02
			> 0,05	Wind–Wiki	0	-0,43	0,43
			< 0,05	Englisch–Wiki[a]	-0,55	-0,95	-0,14
SK02-M	Inhalt	72	**< 0,05**	Wind[a]–Englisch	0,55	0,13	0,98
			> 0,05	Wind–Wiki	0,25	-0,17	0,67
			> 0,05	Englisch–Wiki	-0,30	-0,72	0,11
SK04-M	Suche	72	**< 0,05**	Wind[a]–Englisch	0,62	0,24	1,00
			> 0,05	Wind–Wiki	0,14	-0,30	0,57
			< 0,05[b]	Englisch–Wiki[a]	-0,48	-0,94	-0,03
SK08-M	Suche	72	**< 0,05**	Wind[a]–Englisch	0,59	0,22	0,95
			> 0,05	Wind–Wiki	0,06	-0,38	0,50
			< 0,05[b]	Englisch–Wiki[a]	-0,53	-0,98	-0,08
SK14-M	Suche	72	**< 0,05**	Wind[a]–Englisch	0,64	0,29	0,98
			> 0,05	Wind–Wiki	0,30	-0,11	0,70
			> 0,05	Englisch–Wiki	-0,34	-0,75	0,07
SK16-M	Aufgabe	72	**< 0,05**	Wind[a]–Englisch	0,75	0,31	1,19
			> 0,05	Wind–Wiki	-0,01	-0,38	0,35
			< 0,05[b]	Englisch–Wiki[a]	-0,76	-1,23	-0,30
SK17-M	Suche	72	**< 0,05**	Wind[a]–Englisch	0,50	0,13	0,86
			> 0,05	Wind–Wiki	0,11	-0,29	0,52
			> 0,05	Englisch–Wiki	-0,38	-0,78	0,01
SK-A	Accuracy (EUCS)	72	**< 0,05**	Wind[a]–Englisch	0,59	0,14	1,04
			> 0,05	Wind–Wiki	0,11	-0,34	0,57
			< 0,05[b]	Englisch–Wiki[a]	-0,48	-0,90	-0,06
SK-E-09	EUCS-Skala-2009	72	**< 0,05**	Wind[a]–Englisch	0,47	0,14	0,80
			> 0,05	Wind–Wiki	0,32	-0,03	0,67
			> 0,05	Englisch–Wiki	-0,15	-0,51	0,21
SK-E-13	EUCS-Skala-2013	72	**< 0,05**	Wind[a]–Englisch	0,44	0,15	0,73
			< 0,05[b]	Wind[a]–Wiki	0,27	0,003	0,53
			> 0,05	Englisch–Wiki	-0,17	-0,50	0,15
SK-Z-13	Zusatzskala-2013	72	**< 0,05**	Wind[a]–Englisch	0,51	0,22	0,81
			> 0,05	Wind–Wiki	0,14	-0,23	0,50
			> 0,05	Englisch–Wiki	-0,38	-0,80	0,05
SK-G-13	Gesamtskala-2013	72	**< 0,05**	Wind[a]–Englisch	0,61	0,28	0,94
			> 0,05	Wind–Wiki	0,28	-0,08	0,64
			> 0,05	Englisch–Wiki	-0,33	-0,73	0,07

[a] Bei diesem Thema sind die Probanden zufriedener.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.

Ein besonders auffälliger Befund des dritten Experiments besteht in der Feststellung, dass einige Probanden bei mindestens einer der Aufgaben Suchbegriffe verwenden, die das eigentliche Thema nur am Rande betreffen oder diesem übergeordnet sind (vgl. Abschn. 7.4.1), ein Verhalten, dass auch schon im Kontext des zweiten Experiments zu beobachten ist (vgl. Abschn. 6.4.1). Jedoch erscheint dieses Verhalten im Kontext der Wikiaufgabe, bei der speziell nach Vor- und Nachteilen der Nutzung von Wikis im Schulunterricht gesucht werden soll, besonders problematisch, da nicht ausgeschlossen werden kann, dass Probanden, die den Suchbegriff *wikis* in das Suchfeld eingeben, sich tatsächlich zunächst allgemein über Wikis informieren wollen. In diesem Fall könnte eine große Zahl von Suchergebnissen, die beide Aspekte der Testaufgabe behandeln, zu Irritationen führen und vielleicht sogar die Plausibilität des Untersuchungsdesigns gefährden. Es zeigt sich aber nun, wie im vorangegangenen Absatz beschrieben, dass das Wikithema im Vergleich nicht durchgehend schwerer als die anderen beiden Suchaufgaben ausfällt. Dies kann als erster Hinweis darauf gewertet werden, dass sich die Befürchtung, dieses Thema falle durch das fehlende Hintergrundwissen der Probanden über Wikis aus dem Rahmen, nicht bewahrheitet. Um diese Problematik noch weiter zu analysieren werden neben den soeben beschriebenen Post-Hoc-Tests für SP_A zusätzlich orthogonale Kontraste eingesetzt mit deren Hilfe gezielt untersucht werden soll, ob sich das Wikithema im Schwierigkeitsgrad signifikant von den anderen beiden Themen unterscheidet. Die Details dieser Überprüfung sind ausführlich in Anhang E.5 und speziell in Tabelle E.49 dargestellt. In der überwiegenden Zahl der Fälle (elf von vierzehn Leistungsmaßen) lässt sich kein wenigstens in der Tendenz vorhandener signifikanter Unterschied zwischen dem Wikithema und den beiden anderen Aufgaben nachweisen. Lediglich für die verbleibenden drei Leistungsmaße (M14, V34 u. V35) zeigt die Wikiaufgabe einen höheren Schwierigkeitsgrad. Auch im direkten Vergleich von Wind- und Englischaufgabe bestätigen sich die Ergebnisse von zuvor: In fünf Fällen stellt das Englischthema die leichtere Aufgabe dar, in zwei Fällen die Windaufgabe und für sieben Leistungsmaße lässt sich kein Unterschied feststellen. Insgesamt sprechen diese Befunde dafür, dass das Wikithema keine Sonderrolle einnimmt und somit ohne Einschränkung in die Auswertung mit aufgenommen werden kann.

Bei B10 handelt es sich um das einzige Leistungsmaß, bei dem die Probanden sowohl bei dem Wiki- als auch bei dem Windthema besser als bei der Englischaufgabe abschneiden. Die Testpersonen bewerten also die eher relevanten Dokumente des vermeintlich schwereren Themas signifikant besser. Die Relevanzwahrnehmung bzw. in gewisser Weise die Zufriedenheit mit den Dokumenten fällt also höher aus. Zwar kann aus diesem einzelnen Befund kein allgemeiner Anpassungseffekt postuliert werden, jedoch setzt sich dieser Trend im Kontext der Zufriedenheitsvariablen fort. Für fünfzehn der sechzehn Zufriedenheitsvariablen fallen die Zufriedenheitsurteile der Probanden für das Windthema signifikant höher aus als für das Englischtopic. In einem Fall (F25) hingegen sind die Benutzer mit dem Wikitopic zufriedener. Dies bestätigt insgesamt den zuvor beschriebenen Anpassungseffekt, der zu einer verminderten Zufriedenheit bei leichteren bzw. einer erhöhten Zufriedenheit bei schwereren Aufgaben zu führen scheint. Die im Rahmen der orthogonalen Kontraste erhaltenen Ergebnisse stützen diese Interpretation (vgl. Tab. E.49). Wiederum sind die Probanden mit Ausnahme von F25 jeweils mit der Windaufgabe zufriedener als mit dem Englischtopic. Darüber hinaus ist in keinem Fall ein signifikanter Unterschied zwischen dem Wikithema und den anderen beiden Aufgaben nachweisbar. Somit kann auch im

Tab. 7.27.: Übersicht über beobachtete Topiceffekte in SP_A. Zusätzlich dargestellt sind Topiceffekte, die in SP_B oder unter Ausschluss der Fallgruppen SP_{SB}, SP_{MV}, SP_{TD} und SP_{IZ} eindeutig (E) oder in der Tendenz (T) bestehen bleiben. Der untere Bereich der Tabelle zeigt darüber hinaus die Anzahl der in den betrachteten Teilstichproben zusätzlich hinzukommenden Effekte.

	$\mathbf{SP_A}$	$\mathbf{SP_B}$	$\mathbf{SP_A \setminus SP_{IZ}}$	$\mathbf{SP_A \setminus SP_{MV}}$	$\mathbf{SP_A \setminus SP_{SB}}$	$\mathbf{SP_A \setminus SP_{TD}}$
ID	$n = 24 - 72$	$n = 24 - 24$	$n = 24 - 48$	$n = 24 - 72$	$n = 24 - 48$	$n = 24 - 72$
M14	T	-	-	T	T	-
M35	E	-	-	E	E	T
M36	T	-	-	-	-	E
B10	E	-	-	E	-	T
V09	T	-	-	-	-	-
V34	E	-	-	T	-	E
V35	E	T	-	E	T	T
V37	E	T	-	E	E	E
V38	E	-	-	E	-	T
V51	E	-	E	E	E	E
V52	T	-	-	-	-	-
V54	T	-	-	-	-	-
V79	T	-	-	-	-	-
S01	T	-	-	E	-	T
F06	T	-	-	T	-	E
F17	T	-	-	-	-	-
F18	T	-	E	T	-	T
F22	T	-	-	-	-	T
F25	E	-	-	E	-	T
SK2-M	T	-	-	-	-	-
SK4-M	T	-	-	T	-	E
SK8-M	T	-	-	T	-	T
SK-A	E	-	-	E	-	T
SK14-M	T	-	-	-	-	-
SK16-M	T	-	T	E	-	E
SK17-M	T	-	-	-	-	-
SK-E-09	E	-	-	-	-	T
SK-E-13	E	-	-	T	-	-
SK-Z-13	T	-	-	-	-	T
SK-G-13	E	-	-	-	-	-
	vorhanden	2/14	1/14	9/14	5/14	9/14
BL	zusätzlich	1	0	5	4	5
	gesamt	3	1	14	9	14
	vorhanden	0/16	2/16	8/16	0/16	10/16
BZ	zusätzlich	0	0	4	0	7
	gesamt	0	2	12	0	17

Kontext der Zufriedenheit keine Sonderrolle des Wikithemas festgestellt werden. Gleichzeitig stellt sich die Frage, warum dieser Effekt im zweiten Experiment nicht sichtbar ist. Eine mögliche Erklärung könnte im unterschiedlichen Untersuchungsdesign begründet liegen. Hierbei könnte einerseits das Konstanthalten der Systemleistung und andererseits die erhöhte Anzahl der zu bearbeitenden Aufgaben eine Rolle spielen.

Wie zu Beginn dieses Abschnitts bereits angekündigt, werden die bis hierhin beschriebenen Analysen darüber hinaus noch einmal unter Ausschluss der problematischen Fallgruppen wiederholt. Die entsprechenden Ergebnisse sind detailliert in Anhang E.5 dokumentiert. In Tabelle 7.27 findet sich eine kompakte Übersicht, die im Folgenden erläutert wird. In den Zeilen der Tabelle sind die in Bezug auf die Stichprobe SP_A signifikanten Leistungs- und Zufriedenheitsmaße angegeben. In den Spalten hingegen ist markiert ob die entsprechende Variable auch in Bezug auf die jeweiligen Fallgruppen einen eindeutigen (E), tendenziellen (T) oder keinen (–) Topiceffekt zeigt. Im Fuß der Tabelle kann darüber hinaus nachvollzogen werden, wie viele der Befunde aus SP_A

insgesamt bestätigt werden können bzw. wie viele Topiceffekte für die jeweilige Teilstichprobe neu hinzukommen. Die Stichprobengrößen, die im Kopf der Tabelle angegeben sind, bewegen sich zwischen 24 und 48 Probanden für die Fallgruppen $SP_A \backslash SP_{SB}$ und $SP_A \backslash SP_{IZ}$ und können im Fall von $SP_A \backslash SP_{TD}$ und $SP_A \backslash SP_{MV}$ bis zu 72 Probanden betragen. Die Ergebnisse der am Besten kontrollierten Stichprobe SP_B hingegen basieren durchgehend auf 24 Testpersonen.

Ähnlich wie im zweiten Experiment stellen sich die Topiceffekte in Bezug auf die beiden größten Fallgruppen als weitestgehend stabil heraus. So lassen sich für $SP_A \backslash SP_{MV}$ siebzehn und für $SP_A \backslash SP_{TD}$ neunzehn der 30 in der Hauptauswertung signifikanten Effekte reproduzieren (vgl. Tab. 7.27). Darüber hinaus kommen in der erstgenannten Stichprobe 9 und in der zweiten 12 neue Topiceffekte hinzu. Für die restlichen drei Teilstichproben, $SP_A \backslash SP_{IZ}$, $SP_A \backslash SP_{SB}$ und SP_B hingegen fallen sowohl die Zahl der Übereinstimmungen als auch die Zahl der zusätzlichen Effekte weit geringer aus. Dies legt den Schluss nahe, dass sich die Teilstichproben $SP_A \backslash SP_{MV}$ und $SP_A \backslash SP_{TD}$ in Bezug auf den Topiceffekt nur wenig von der Stichprobe SP_A unterscheiden. Eine nicht geglückte Erwartungsmanipulation oder ein Durchschauen des Testdesigns scheinen also nur einen geringen Einfluss auf das Auftreten von Topiceffekten zu haben, da ein Ausschluss dieser Gruppen insgesamt nicht zu einer merklichen Reduktion dieser Effekte führt. Vielmehr kann nun auch für weitere Variablen ein Topiceffekt nachgewiesen werden, was höchstwahrscheinlich auf die bessere Kontrolliertheit dieser Stichproben zurückzuführen ist. Darüber hinaus bestätigen die Post-hoc-Tests (vgl. Tab. E.51 bis Tab. E.58) auch inhaltlich die Übereinstimmung mit SP_A. So erweist sich das Englischthema für den überwiegenden Teil der Variablen wiederum als leichtere Aufgabe, während in Bezug auf andere Benutzerleistungsaspekte das Wind- bzw. das Wikithema leichter zu bearbeiten sind. Dies gilt außerdem für die zuvor beschriebenen Befunde in Bezug auf die Benutzerzufriedenheit.

Der starke Rückgang der Topiceffekte in Bezug auf die Stichproben $SP_A \backslash SP_{SB}$, $SP_A \backslash SP_{IZ}$ und SP_B spricht hingegen dafür, dass der größte Beitrag zu den Topiceffekten von den Fallgruppen SP_{SB} und SP_{IZ} ausgeht. Sobald nämlich diese Gruppen aus der Stichprobe SP_A ausgeschlossen werden, verschwindet die überwiegende Mehrheit der zuvor in SP_A signifikanten Topiceffekte. Dabei ist zu beachten, dass an dieser Stelle aufgrund der Gruppengröße im Gegensatz zur zweiten Nutzerstudie die Gruppe SP_{IZ} anstelle von SP_{UZ} aus der Stichprobe SP_A entfernt wird. Die Tatsache, dass sich die Ergebnisse tatsächlich spiegelbildlich verhalten, der Ausschluss von SP_{IZ} führt zu einem Rückgang der Topiceffekte, wohingegen in Experiment 2 der Ausschluss von SP_{UZ} nur einen geringen Effekt ausübt, spricht noch einmal für die Konsistenz der Ergebnisse. Allerdings sind die verbleibenden Stichprobenumfänge im direkten Vergleich zu Experiment 2 groß genug, um diesen Rückgang nicht alleine durch eine zu geringe Anzahl an Testpersonen zu erklären. Vielmehr kann diese Beobachtung als Hinweis darauf gewertet werden, dass der Topiceffekt zum einen in der Teilnehmergruppe mit unspezifischen Suchbegriffen (SP_{SB}) und zum anderen in der Teilnehmergruppe die bei allen Suchaufgaben die volle Suchzeit ausschöpfen (SP_{IZ}) zu verorten ist. Da beide Teilstichproben jeweils 33 Testpersonen umfassen, aber insgesamt nur sieben Probanden beiden Gruppen zugeordnet werden können, ist weiterhin davon auszugehen, dass beide Fallgruppen unabhängig voneinander Topiceffekte verursachen. In Bezug auf die Gruppe SP_{SB} scheint dies direkt plausibel, da die Fähigkeit zu einem Thema eine spezifische Suchanfrage stellen zu können auch mit der Relevanzwahrnehmung der Dokumente

verknüpft ist. Für die Fallgruppe SP_{IZ} hingegen ist folgende Interpretation denkbar. Teilnehmer, die bei ihrer Suchaufgabe selbstbestimmt abbrechen, bearbeiten eine Aufgabe bis sie mit dem Resultat ihrer Suche zufrieden sind. Dies führt im Umkehrschluss zu vergleichbaren Ergebnissen für jede der Suchaufgaben unabhängig vom bearbeiteten Topic. Sucht ein Teilnehmer hingegen bei jeder Aufgabe für genau zehn Minuten, so wird er je nach Schwierigkeitsgrad der Aufgabe eine bessere oder schlechtere Leistung erbringen. Aus diesem Grund erscheint die Reduzierung des Topiceffekts in Bezug auf die Fallgruppe SP_{IZ} ebenfalls plausibel. Dessen unbenommen ist zu beachten, dass sowohl $SP_A \backslash SP_{SB}$ als auch $SP_A \backslash SP_{IZ}$ über geringere Fallzahlen verfügen und somit nicht ausgeschlossen werden kann, dass dies zusätzlich zu einer Verringerung signifikanter Effekte führt. Für die verbleibenden Topiceffekte ergeben die Post-hoc-Tests (vgl. Tab. E.51 bis Tab. E.58) hingegen analoge Resultate zur Stichprobe SP_A. Es zeigt sich also wiederum, dass anders als zuvor vermutet, die Topiceffekte weniger im Zusammenhang mit der Wikiaufgabe als vielmehr in Bezug auf das Englischtopic auftreten. Somit kann selbst für die Fallgruppe SP_{SB} eine Sonderrolle des Wikithemas ausgeschlossen werden, obwohl unspezifische Suchen gerade im Kontext dieser Suchaufgabe zu beobachten sind (vgl. Abschn. 7.4.1).

Zusammenfassend kann somit festgestellt werden, dass in der Stichprobe SP_A ähnlich wie im zweiten Experiment Topiceffekte nachweisbar sind und somit, wie für die Hauptauswertung geschehen, ein topicbezogenes Balancieren der Untersuchungsgruppen geraten scheint. Der unterschiedliche Schwierigkeitsgrad der einzelnen Aufgaben trägt auf der anderen Seite jedoch auch zum Realitätsgrad der Untersuchung bei. Darüber hinaus kann das Auftreten von Topiceffekten in zwei der als problematisch eingestuften Fallgruppen (SP_{SB} und SP_{IZ}) verortet werden. Die Befürchtung, dass die Wikiaufgabe durch fehlendes Hintergrundwissen eine Sonderrolle einnimmt scheint sich hingegen nicht zu bewahrheiten. In den meisten Fällen in denen Topiceffekte auftreten, sind keine signifikanten Unterschiede zwischen dem Wiki- und den anderen beiden Suchthemen nachweisbar. Darüber hinaus kann eine interessante positive Korrelation zwischen Aufgabenschwierigkeit und Benutzerzufriedenheit festgestellt werden, die allerdings eingehender untersucht werden müsste.

7.4.6.2. Personenbezogene Störfaktoren

In diesem Abschnitt werden die Befunde der Hauptanalyse mit Blick auf den Einfluss personenbezogener Störfaktoren untersucht. Dazu wird im dritten Experiment zunächst folgende Gruppe von Kovariaten berücksichtigt: Alter (K01), Geschlecht (K02), Muttersprache (K03) sowie fünf Erfahrungsvariablen (Computernutzungsjahre (K04), Computernutzungsstunden (K05), Selbsteinschschätzung Suchmaschinenwissen (K08), Suchmaschinennutzungsjahre (K09) u. Suchmaschinennutzungsstunden (K10)). Da diese Kovariaten für jede Testperson erhoben werden, können die entsprechenden Kovarianzanalysen auf Grundlage der bereits im Rahmen der Hauptanalyse verwendeten Zufallsstichproben durchgeführt werden. Weiterhin ist darauf hinzuweisen, dass aus untersuchungsökonomischen Gründen einige der im zweiten Experiment betrachteten Kovariaten entfallen, die Benennung der Kovariaten jedoch zu Vergleichszwecken beibehalten wird (vgl. Abschn. 7.3.3). Im Rahmen der Zufriedenheitsauswertung werden darüber hinaus eine Reihe leistungsbezogener Kovariaten untersucht, um herauszufinden, inwiefern die Wahrnehmung der selbst erbrachten Suchleistung das Zufriedenheitsurteil beeinflusst. Dabei handelt es sich um: die Zeit bis zum ersten richtig relevant bewerteten Dokument bei binärer bzw. 4-

stufiger Relevanzskala (S05 u. S06), die Menge relevant bewerteter Dokumente bei binärer bzw. 4-stufiger Relevanzskala (M10 u. M37), die Menge richtig relevant bewerteter Dokumente bei binärer bzw. 4-stufiger Relevanzskala (M16 u. M45), die Anzahl der Suchen (S01), die erste betrachtete Rankingposition (S02), die letzte betrachtete Rankingposition (S03) und die Suchdauer (S04). Im Kontext der wahrgenommen Suchmaschinenqualität werden darüber hinaus die folgenden Leistungsmaße betrachtet: Die durchschnittliche Bewertung relevanter Dokumente insgesamt (B04 u. B16), sowie in Bezug auf die erste (B05 u. B17) und letzte Suche (B06 u. B18) sowie die entsprechenden mittleren Bewertungen der eher relevanten Dokumente (B10, B11, B12), der eher irrelevanten Dokumente (B07, B08, B09) und der irrelevanten Dokumente (B01 u. B13, B02 u. B14, B03 u. B10). Da diese Leistungsmaße nicht zwingenderweise für alle Testpersonen vorhanden sind, müssen in diesem Fall im Gegensatz zur Auswertung der demographischen und erfahrungsbezogenen Kovariaten jeweils neue Zufallsstichproben generiert werden.

Wie bereits im Kontext von Experiment 2 erläutert (vgl. Abschn. 6.4.6.2), führt das angewendete Auswertungskonzept dazu, dass im Rahmen der Überprüfung personenbezogener Störfaktoren sehr viele Analysen berechnet werden müssen. Als Datengrundlage dienen dabei wiederum die über alle drei Aufgaben gemittelten Variablenwerte der Gesamtstichprobe SP_A. Von einer Darstellung der bereinigten Stichprobe SP_B wird hingegen aus Platzgründen abgesehen. So ergeben sich im ersten Analyseschritt ausgehend von den 115 Leistungsmaßen in SP_A mit jeweils fünf Zufallsstichproben und den 8 demographischen und erfahrungsbezogenen Kovariaten (K01 bis K10) im Höchstfall 9.200 Analysen, wenn klassische und robuste ANCOVA-Berechnungen berücksichtigt werden. Da für die 54 in SP_A enthaltenen Zufriedenheitsindikatoren zusätzlich die 28 leistungsbezogenen Kovariaten analysiert werden, ergeben sich hier im Höchstfall 19.440 Auswertungen.

Wie im zweiten Experiment werden deshalb aufgrund der Fülle der Daten im Folgenden ausschließlich Kovarianzanalysen betrachtet, die alle statistischen Voraussetzungen erfüllen und darüber hinaus eine zumindest tendenzielle Aussage über alle fünf Zufallsstichproben zulassen. Dabei setzt die Durchführung einer Kovarianzanalyse neben der Normalverteilung und Varianzhomogenität der Daten, die Unabhängigkeit der Kovariaten vom Behandlungseffekt sowie die Homogenität der Regressionssteigungen voraus (vgl. Abschn. 4.3.2.3). Des Weiteren sollte die Kovariate nicht über alle Testpersonen hinweg konstant sein. Während bei Verletzung der Normalverteilung oder der Varianzhomogenität robuste Analysen durchgeführt werden können, führen Verletzungen der zusätzlichen Voraussetzungen zu einem Ausschluss der korrespondierenden Variablen-Kovariaten-Kombination für die personenbezogenen Kovariaten. Die Natur der leistungsbezogenen Kovariaten als direkte Reaktion der Testpersonen auf das experimentelle Treatment hingegen führt in natürlicher Weise zu Abhängigkeiten von den unabhängigen Variablen. Da in diesem Fall aber gerade der Einfluss des eigenen Sucherfolgs auf das Zufriedenheitsurteil herauspartialisiert werden soll, führt eine solche Abhängigkeit nicht zu einem Ausschluss der betreffenden Kovariate. Vor dem Hintergrund der Tatsache, dass solche Korrelationen lediglich zu einer Reduzierung signifikanter Effekte führen, erscheint dieses Vorgehen jedoch unproblematisch (vgl. Abschn. 4.3.2.3). Weiterhin führen numerische Instabilitäten, vermutlich aufgrund geringer Gruppengrößen, in einigen Fällen dazu, dass die geschätzten Mittelwerte außerhalb der zu erwartenden Wertebereiche liegen. Man erhält bspw. negative Werte obwohl alle Antworten

der Testpersonen zwischen Null und Eins liegen. Am Häufigsten passiert dies für die Kovariate Alter (K01). Auch diese offensichtlich fehlerhaften Analysen werden von der weiteren Auswertung ausgenommen.

Danach verbleiben für die demographischen und erfahrungsbezogenen Kovariaten noch 12.150 Tests zur Auswertung, die sich je nachdem ob die robuste oder klassische Analyse zu werten ist, schlussendlich auf 6.075 Analysen reduzieren. Im Fall der leistungsbezogenen Kovariaten führt, wie bereits erläutert, eine Abhängigkeit zwischen Kovariate und Zufriedenheitsindikator nicht zu einem Ausschluss der Analyse, so dass hier 9.170 bzw. nach Wahl der entsprechenden klassischen oder robusten Analyse 4.585 Tests verbleiben. Ausgewählte Ergebnisse sind in den Tabellen 7.28, 7.29 und 7.30 zusammengefasst. Detailliertere Informationen, wie die entsprechenden Mittelwerte und Signifikanzniveaus, können den Tabellen in den Abschnitten E.6.2 und E.6.3 im Anhang entnommen werden. Weiterhin geben die Tabellen E.59 und E.60 in Anhang E.6.1 eine Übersicht über die Variablen, für welche sich keine Änderungen der Effekte durch die Kovarianzanalyse ergeben, wobei dieser Fall im Zusammenhang mit den leistungsbezogenen Kovariaten nicht auftritt.

Die in den Tabellen 7.28 bis 7.31 dargestellten Ergebnisse sind getrennt nach Leistungsmaßen und Zufriedenheitsindikatoren aufgeteilt. Die Zeilen enthalten die abhängigen Variablen, wohingegen die Spalten den getesteten Kovariaten entsprechen. Jede Zelle informiert über das Hinzukommen (+) oder Verschwinden (−) eines Haupteffekts (S = System, E = Erwartung) oder einer Interaktion (I). So steht das Kürzel $+I_{SE}$ bspw. für einen hinzukommenden Interaktionseffekt zwischen Systemleistung und Erwartungshaltung. Im Folgenden werden zunächst die Ergebnisse der Kovarianzanalyse in Bezug auf demographische und erfahrungsbezogene Kovariaten dargestellt.

Demographische und erfahrungsbezogene Kovariaten

Wie in Experiment 2 bestätigen sich weitgehend die schon in der Hauptuntersuchung diskutierten Befunde. Im Kontext der Benutzerleistung enfallen sieben signfikante Effekte, wohingegen fünf Haupt- und zwei Interaktionseffekte neu hinzukommen. Für die Zufriedenheit verliert unter Berücksichtigung demographischer und erfahrungsbezogener Kovariaten lediglich ein Haupteffekt seinen signifikanten Einfluss, während auf der anderen Seite vier Haupteffekte und eine Interaktion zusätzlich signifikant werden.

Als erstes werden die Ergebnisse im Kontext der Benutzerleistung besprochen. Hinsichtlich der sieben verschwindenden Effekte (M02, M26, B06, Z02, Z05-log, V14 u. V58) ist zunächst darauf hinzuweisen, dass nur zwei der Variablen (M02 u. Z05-log) im Rahmen der Hauptuntersuchung in beiden Stichproben SP_A und SP_B signifikante Effekte zeigen. Diese Instabilität spiegelt sich nun also auch in Bezug auf die Kovarianzanalyse wider. Obwohl mit B06 und M26 zwei Indikatoren für den systemabhängigen und mit V14 ein Indikator für den erwartungsabhängigen Anpassungseffekt der Relevanzbewertung entfallen, zeigt die Kovarianzanalyse insgesamt gesehen dennoch, dass die meisten der Variablen, die diesen Befunde zugrunde liegen, weiterhin signifikant bleiben. Genauer, kann Tabelle E.59 entnommen werden, dass bspw. weiterhin die Variablen M07, M14, V05, V06, V13, V40 und V44 den systembedingten Anpassungseffekt stützen. Gleiches gilt für die Variablen V12, V11 und V08 in Bezug auf den erwartungsbedingten Anpassungseffekt. Darüber hinaus zeigt sich die im Rahmen der Hauptauswertung sowohl für SP_A als auch SP_B signifikante

Tab. 7.28.: Übersicht über Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses demographischer und erfahrungsbezogener Störfaktoren auf die Benutzerleistung in SP_A. Die Tabelle stellt für jede Kovariate alle Leistungsmaße mit entfallenden (–) bzw. hinzukommenden (+) Effekten der Systemleistung (S), der Erwartung (E) oder ihre Interaktion (I) dar.

ID	Beschreibung	K02	K03	K04	K05	K08	K09	K10
M02	Anz. aufg. Dok. (erste 10 Dok.)	$-I_{SE}$	$-I_{SE}$	$-I_{SE}$	-	$-I_{SE}$	$-I_{SE}$	-
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	-	$+I_{SE}$	-	-	-	$+I_{SE}$	-
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	$+I_{SE}$	-	-	-	+S	-	-
M21	Anz. aufg. eher rel. Dok. (4-st.)	-	-	-	-	-	+S	+S
M26	Anz. falsch eher irrel. bew. Dok. (4-st.)	−S	-	-	−S	−S	-	-
M29	Anz. falsch eher irrel. bew. rel. Dok. (4-st.)	-	+E	+E	-	-	+E	+E
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	-	-	-	-	-	-	−S
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	-	-	-	−S	−S	-	−S
Z05-log	Durchschn. Betrachtungsz. irrel. bew. Dok.	−S	-	-	-	-	-	−S
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	-	-	−E	-	-	-	-
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	+S	+S	+S	+S	+S	+S	+S
V34	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$ (4-st.)	-	-	-	+S	-	+S	+S
V58	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$ (4-st.)	−E	−E	-	-	-	-	-

Tab. 7.29.: Übersicht über Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses demographischer und erfahrungsbezogener Störfaktoren auf die Benutzerzufriedenheit in SP_A. Die Tabelle stellt für jede Kovariate alle Zufriedenheitsindikatoren mit entfallenden (–) bzw. hinzukommenden (+) Effekten der Systemleistung (S), der Erwartung (E) oder ihre Interaktion (I) dar.

ID	Beschreibung	K02	K03	K04	K05	K09	K10
F11	Liefert die Suchmaschine aktuelle Information?	-	+E	-	-	-	-
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	−S	-	-	-	-	-
SK03-M	Benutzerfreundlichkeit	-	-	-	$+I_{SE}$	-	-
SK14-M	Suche	+S	-	-	+S	-	-
SK15-F	Eigenleistung	+E	+E	-	+E	+E	+E
E01	Ich glaube, ich werde in zehn Minuten ... relevante Dokumente finden.	-	-	+E	-	-	+E

Interaktion für das Leistungsmaß M02 als relativ instabil bezüglich der Kovarianzanalyse, da dieser Effekt für insgesamt fünf unterschiedliche Kovariaten verschwindet (K02, K03, K04, K08 u. K09).

Mit dem Wegfall der signifikanten Systemeffekte für die Variablen Z02-log und Z05 zeigt sich einmal mehr, dass Unterschiede in Bezug auf Bearbeitungszeiten schwer zu erfassen sind. Es erscheint jedoch einleuchtend, dass eine andere Muttersprache oder geringere Computererfahrung zu einer erhöhten Bearbeitungszeit führen. Besonders interessant erscheinen dabei der Einfluss der Muttersprache und des Geschlechts, die in Folgestudien weiter untersucht werden sollten.

Die im Rahmen der Kovarianzanalyse neu auftretenden Effekte lassen sich grob in zwei Gruppen unterteilen. Die erste Gruppe umfasst abhängige Variablen, die zuvor ausschließlich in SP_B signifikante Effekte zeigen, die nun ebenfalls in SP_A nachweisbar sind (V29 u. M17). Ihr Einfluss

wird durch die Kovarianzanalyse nun also auch in der weniger kontrollierten Stichprobe SP_A sichtbar. In die zweite Gruppe hingegen fallen Variablen für die ein Einfluss der unabhängigen Variablen neu hinzukommt (M11, M21, M29 u. V34). Dabei bestätigt M29 erneut den bereits bekannten erwartungsbedingten Anpassungseffekt. M21 und V34 hingegen zeigen, dass Nutzer des besseren Systems mehr eher relevante Dokumente aufrufen als Nutzer des schlechteren Systems. Die entsprechenden Mittelwerte können in Tabelle E.61 in Anhang E.6.2 nachvollzogen werden. Dieser Effekt an sich ist nicht überraschend, da die Ergebnislisten des besseren Systems mehr eher relevante Dokumente enthalten, als die des schlechteren Systems. Interessant hingegen ist die Tatsache, dass dieser Effekt erst dann deutlich wird, wenn man die Vorerfahrung der Testpersonen mit Suchmaschinen mit einbezieht. Diese Kovariaten sollten somit in Folgestudien immer mit erhoben werden. Abschließend bleibt noch auf die neu hinzukommende Wechselwirkung für M11 einzugehen. Hier zeigt sich, wie Abbildung 7.22 zu entnehmen, eine gegenseitige Verstärkung von gutem System und hoher Erwartungshaltung: Teilnehmer in dieser Untersuchungsgruppe bewerten eine größere Anzahl der ersten zehn Dokumente als relevant als alle übrigen Teilnehmer. Der Umstand, dass im Fall des besseren Systems mehr relevante Dokumente enthalten sind, wird von Testpersonen mit der hohen Erwartungshaltung also stärker wahrgenommen. Die Tatsache, dass diese Beobachtung im Widerspruch zu dem systembedingten Anpassungseffekt zu stehen scheint, kann als weiterer Hinweis darauf gesehen werden, dass dieser Effekt, wie schon im Rahmen der dynamischen Auswertung der Benutzerleistung vermutet (vgl. Abb. 7.15 in Abschn. 7.4.3.3), einer zusätzlichen Dynamik innerhalb des Suchprozesses unterworfen ist.

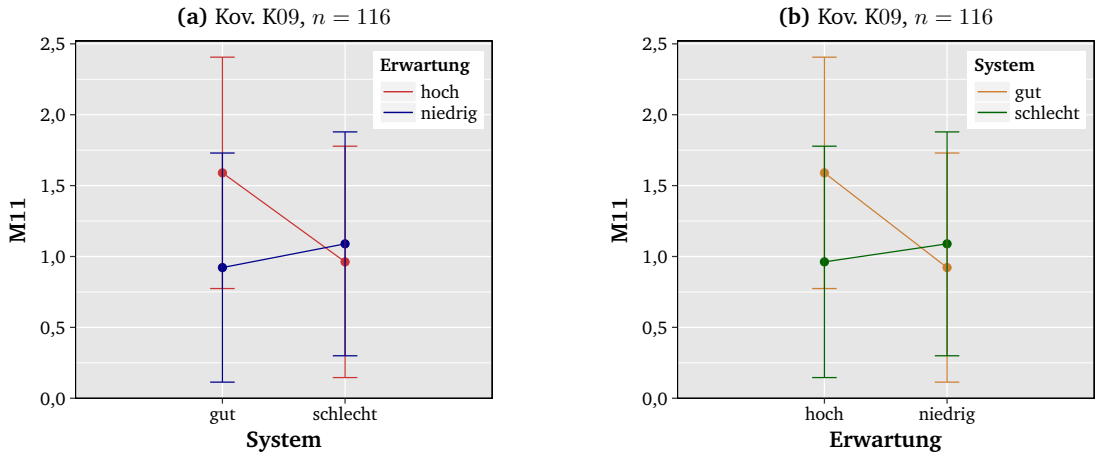


Abb. 7.22.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für die Anzahl der ersten zehn angezeigten Dokumente, die als relevant bewertet werden (M11) unter Berücksichtigung der Suchmaschinennutzungsjahre (K09) als Kovariate. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Ein Unterschied zwischen den beiden Systemgüten ist nur im Kontext der hohen Erwartungshaltung erkennbar (b). Die größere Anzahl relevanter Dokumente im Kontext der besseren Systemleistung wird somit nur von Probanden mit einer höheren Erwartungshaltung wahrgenommen. Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte.

Als nächstes werden die Ergebnisse im Kontext der Benutzerzufriedenheit diskutiert und zunächst auf die personenbezogenen Kovariaten K01 bis K10 eingegangen. Wie bereits erwähnt,

verliert diesbezüglich im Rahmen der Zufriedenheitsauswertung lediglich ein Haupteffekt (F19) seinen signifikanten Einfluss, während auf der anderen Seite vier System- und Erwartungseffekte (F18, SK14-M, F11, SK15-F) sowie eine Interaktion (SK3-M) hinzukommen. Bemerkenswert ist in diesem Zusammenhang, dass mit F11 und SK15-F nun auch die letzten beiden Variablen ohne einen in SP_A abgesicherten Behandlungseffekt (vgl. Abschn. 7.4.5) zeigen, dass die Benutzererwartung eine entscheidende Rolle bei der Zufriedenheitsbeurteilung spielt. In Bezug auf die Richtung dieser Effekte lassen sich die Befunde der Hauptauswertung wiederum bestätigen. So führt erneut sowohl eine höhere Erwartungshaltung der Probanden als auch eine bessere Systemqualität in allen Fällen zu einer höheren Zufriedenheit. Gleichermaßen stützt der zusätzliche Erwartungseffekt im Zusammenhang mit der von den Probanden für die Folgeaufgabe geschätzten Anzahl gefundener relevanter Dokumente (E01) den in Abschnitt 7.4.5 dargelegten Befund einer positiven Abhängigkeit von der induzierten Erwartungshaltung.

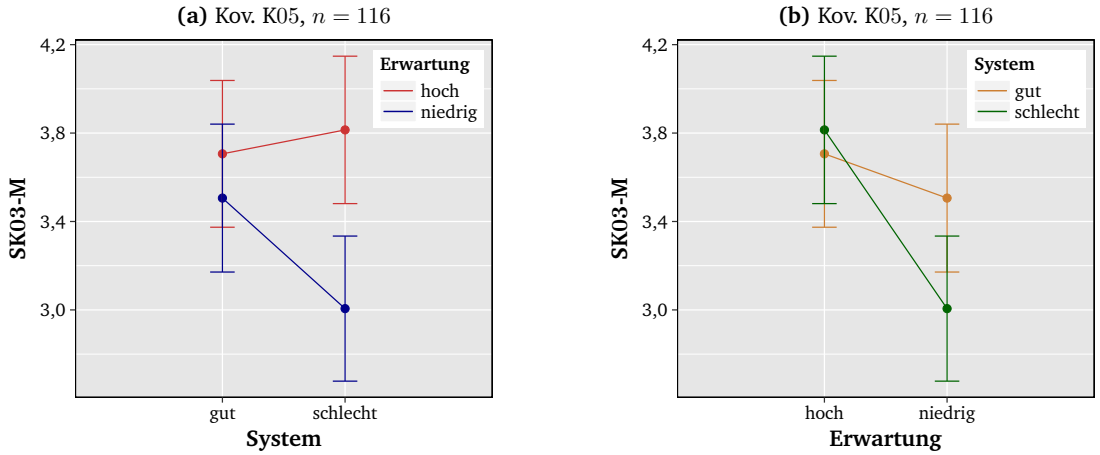


Abb. 7.23.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für die Benutzerfreundlichkeit (SK03-M) unter Berücksichtigung der Computernutzungsstunden (K05) als Kovariate. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bild (a): Die hohe Erwartungshaltung führt über beide Systemgüten hinweg tendenziell zu einer höheren Zufriedenheit im Vergleich zur niedrigen Erwartungshaltung. Während sich für die hohe Erwartungshaltung darüber hinaus keine Systemabhängigkeit zeigt, fällt die Zufriedenheit mit dem schlechteren System bei der niedrigen Erwartungshaltung gegenüber dem besseren System ab. Bild (b): Eine Abhängigkeit von der Systemleistung ist für die Zufriedenheit mit der Benutzerfreundlichkeit nur im Kontext der niedrigen Erwartungshaltung erkennbar. Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte.

Während die Einbeziehung der Kovariaten auf Ebene der Haupteffekte also zu keinen wesentlichen Änderungen der Ergebnisse führt, ist die hinzukommende Wechselwirkung umso interessanter. Nachdem nämlich der Einfluss der Computererfahrung (K05) aus der abhängigen Variable SK03-M herauspartialisiert wird, bestätigen sich hier, wie in Abbildung 7.23 zu sehen, zumindest teilweise die Vorhersagen des C/D-Paradigmas. So lässt sich ablesen, dass der Systemunterschied bei hoher Erwartungsmanipulation ungefähr gleich ausfällt, die Testteilnehmer bei niedriger Erwartungshaltung jedoch mit dem besseren System zufriedener sind (positive Erwartungsdiskonfirmation). Ein Einfluss der Systemleistung lässt sich hingegen nur im Kontext der niedrigen Erwartungshaltung feststellen (Bild(b)). Dies ist konsistent mit der Beobachtung in Ab-

schnitt 7.4.3.2, dass der den wahrgenommenen Systemunterschied reduzierende systembedingte Anpassungseffekt der Relevanzwahrnehmung eng mit einer hohen Erwartungshaltung verknüpft ist (vgl. Abb. 7.12). Die Tatsache, dass die Probanden unabhängig von der Systemleistung mit einer höheren Erwartungshaltung zufriedener sind, legt erneut die Vermutung nahe, dass die Wahrnehmung der eigenen Suchleistung aufgrund es erwartungsbedingten Anpassungseffekts der Relevanzwahrnehmung ebenfalls eine Rolle spielt. Es erscheint deshalb geboten in einem weitergehenden Analyseschritt die erbrachte Eigenleistung der Testpersonen zu berücksichtigen um auf diese Weise den komplexen inneren Prozessen bei der Zufriedenheitsbeurteilung Rechnung zu tragen. Die entsprechenden Ergebnisse werden im Folgenden beschrieben.

Leistungsbezogene Kovariaten

Eine weitere Frage im Kontext der Auswertung der Benutzerzufriedenheit betrifft die Abhängigkeit des Zufriedenheitsurteils von der Wahrnehmung der eigenen Suchleistung durch die Testpersonen. Wie eingangs beschrieben, erscheint es plausibel, dass die Eigenleistung den Einfluss von Systemgüte und Erwartungshaltung auf die Zufriedenheit der Teilnehmer überlagern könnte. Aus diesem Grund wird im Folgenden eine Kovarianzanalyse für die Zufriedenheitsindikatoren bezüglich einiger Leistungsmaße durchgeführt, die ausgewählte Aspekte des Aufwandes, der Effektivität sowie der Relevanzwahrnehmung der Testpersonen widerspiegeln (vgl. Tab. D.7 in Anh. D.4). Da die zugrundeliegenden Leistungsmaße naturgemäß nicht für alle Testpersonen vorliegen, müssen im Gegensatz zu den im vorigen Abschnitt beschriebenen Kovarianzanalysen für jede Variablen-Kovariaten-Kombination fünf neue Zufallsstichproben gezogen werden. Dies stellt sicher, dass sich die Auswertung ausschließlich auf vollständige Fälle stützt, die keine fehlenden Werte enthalten. Um möglichst viele Leistungsmaße überprüfen zu können wird die Mindeststichprobengröße auf sechs Teilnehmer pro Treatmentgruppe herabgesetzt. Die Leistungskovariaten B02, B07, B08, B09, B13, B14 sowie B15 erfüllen jedoch für keine der Zufriedenheitsvariablen diese Mindestvoraussetzung und sind aus diesem Grund in den folgenden Tabellen nicht enthalten. Es handelt sich dabei neben der mittleren Bewertung irrelevanter Dokumente bei der ersten Suche um alle Bewertungen, die sich auf die eher irrelevanten oder irrelevanten Dokumente im Kontext der 4-stufigen Relevanzskala beziehen. Die Resultate für die übrigen Leistungskovariaten sind in den Tabellen 7.30 und 7.31 zusammengefasst. Ihr Aufbau entspricht der bereits am Ende des Abschnitts 7.4.6.2 erläuterten Struktur, bei der hinzukommende und verschwindende signifikante Haupteffekte und Interaktionen jeweils durch ein vorangestelltes $+$ bzw. $-$ ausgewiesen werden.

Zunächst fällt auf, dass sich die These zu bestätigen scheint, dass die Zufriedenheit der Testpersonen eng mit der selbst erbrachten Suchleistung zusammenhängt: Von den insgesamt 54 Zufriedenheitsindikatoren zeigen 51 einen Wegfall bzw. ein Hinzukommen von Effekten. Gleiches gilt auch für die fünf erwartungsbezogenen Frageitems (E01 bis E05) sowie deren zugehörige Mittelwertskala (E06-M). Einzig die Frageitems F09 und F11, sowie der Mittelwert über die Zusatzitems SK-Z-13 bleiben ungeändert, wenn die Eigenleistung der Probanden mit in die Auswertung einbezogen wird. Dies erscheint insbesondere für die beiden Frageitems plausibel, da hier die Aktualität der Ergebnisse bzw. die Einfachheit der Bedienung abgefragt wird, die nicht unmittelbar an die eigene Suchleistung gekoppelt sind. Der Großteil der Änderungen lässt sich in drei Gruppen unterteilen, je nachdem, ob es zu einem Wegfall des Erwartungseinflusses, einem

Tab. 7.30.: Übersicht über Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses leistungsbezogener die Relevanzwahrnehmung betreffender Störfaktoren auf die Benutzerzufriedenheit in SP_A. Die Tabelle stellt für jede Kovariate alle Zufriedenheitsindikatoren mit entfallenden (–) bzw. hinzukommenden (+) Effekten der Systemleistung (S), der Erwartung (E) oder ihre Interaktion (I) dar.

ID	Beschreibung	B01	B03	B04	B05	B06	B10	B11	B12	B16	B17	B18
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	-	-	-	-	-	-	−E	-	-	+S	-
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	−E	-	+S[d]	+S	-	-	-	-	+S[d]	+S	-
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	-	-	+S[d]	-	-	−E[d]	−E	-	+S[d]	-	-
F04	Liefert die Suchmaschine genügend Information?	-	-	-	-	-	−E[d]	−E	-	-	-	-
F05	Ist die Suchmaschine präzise?	-	-	+S[d]	-	+S[d]	-	-	+S	+S[d]	-	-
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	-	-	-	-	-	-	-	+S	+S[d]	-	-
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	-	-	-	−E	-	−E[d]	−E	-	-	-	-
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	-	-	-	-	-	−E[d]	−E	−E	-	-	-
F12	Ist die Suchmaschine erfolgreich?	-	-	-	-	-	-	-	-	-	-	-
F13	Sind Sie mit der Suchmaschine zufrieden?	-	-	-	-	+S[d]	-	−E	-	+S[d]	-	-
F14	Es war einfach, die Aufgabe zu bearbeiten.	-	-	-	-	−E[d]	-	−E	−E	-	−E	−E[d]
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	-	-	-	+S	-	-	−E	-	-	-	-
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	-	-	-	−E	+S[d]	−E	-	+S −E	+S[d]	−E	−E
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	-	-	+S[d]	+S	+S[d]	-	-	+S	+S[d]	-	+S −E
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	-	-	-	−S	-	−S	-	−S −E	-	−E	-
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	-	-	-	−E	-	−E	-	-	-	−E	−E
F22	Ich bin mit den Suchergebnissen zufrieden.	-	-	+S[d]	+S	-	-	+S −E	+S	+S[d]	-	-
F23	Ich bin mit meiner Suchleistung zufrieden.	-	-	-	−E	−E[d]	−E	−E	-	−E[d]	−E	−E
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	-	-	-	+S −E	-	-	−E	+S	+S[d]	+S	+S[d]
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	−E	-	-	-	-	-	−E	-	+S[d]	$+I_{SE}$	-
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	-	−E	-	-	-	-	-	-	+S[d]	−E	-
SK01-M	Genauigkeit	-	-	-	-	-	-	−E	-	+S[d]	-	-
SK02-M	Inhalt	−E	-	-	+S	-	-	−E	-	-	-	-
SK03-M	Benutzerfreundlichkeit	-	-	+S[d]	-	-	-	-	-	+S[d]	$+I_{SE}$	$+I_{SE}$ +S
SK04-M	Suche	-	-	+S[d]	+S	+S[d]	-	-	+S	+S[d]	+S	+S
SK07-M[b]	Benutzerfreundlichkeit	−E	-	+S[d]	-	+S[d]	-	-	+S	+S[d]	+S	+S

[a] Entspricht auch den Skalen SK15-M und SK19-M.

Fortsetzung auf nächster Seite

[b] Entspricht auch der Skala SK18-M.

[c] Entspricht auch der Skala SK13-M.

[d] Es besteht eine Abhängigkeit zwischen Kovariate und abhängiger Variable.

[e] Teilweise ergeben sich negative Gruppenmittelwerte.

Tab. 7.30 (Fortsetzung)

ID	Beschreibung	B01	B03	B04	B05	B06	B10	B11	B12	B16	B17	B18
SK08-M	Suche	-	-	-	-	-	-	-	+S	+S^{d}	-	+S
SK09-M	Benutzerfreundlichkeit	-	-	+S^{d}	-	+S^{d}	-	-	+S	+S^{d}	+I$_{SE}$ +S	+S
SK11-M[a]	Eigenleistung	-	-	-	-	−E^{d}	−E^{d}	-	-	-	−E	−E
SK12-F	Suchergebnis	-	-	-	-	-	-	−E	-	+S^{d}	-	-
SK12-M	Suchergebnis	-	-	-	-	-	-	-	+S	-	-	-
SK13-F	Benutzerfreundlichkeit	−E	-	-	-	-	-	-	-	-	+I$_{SE}$	-
SK14-F	Suche	-	-	+S^{d}	+S −E	+S^{d}	-	−E	+S	+S^{d}	-	+S
SK14-M	Suche	-	-	+S^{d}	+S	+S^{d}	+S^{d}	-	+S	+S^{d}	+S −E	+S
SK15-F	Eigenleistung	-	+E	+E^{d}	+E	-	-	-	-	-	-	-
SK16-F	Aufgabe	−E	−E	-	−E	−E^{d}	-	−E	+S −E	-	-	−E
SK16-M	Aufgabe	-	-	-	-	-	-	-	+S	+S^{d}	-	-
SK17-F	Suche	-	-	-	-	-	-	−E	-	+S^{d}	-	-
SK17-M	Suche	-	-	-	-	+S^{d}	-	-	-	-	-	+S
SK18-F	Benutzerfreundlichkeit	-	−E	+S^{d}	+S −E	+S^{d}	+S^{d}	−E	+S	+S^{d}	−E	+S
SK19-F	Eigenleistung	-	-	-	−E	−E^{d}	−E^{d}	-	-	−E^{d}	−E	−E
SK-A	Accuracy (EUCS)	−E	-	-	-	-	-	+S	+S	+S^{d}	+S	+S
SK-C	Content (EUCS)	-	-	-	+S	-	-	−E	-	-	-	-
SK-E[c]	Ease of Use (EUCS)	−E	-	-	-	-	-	-	-	-	+I$_{SE}$	-
SK-T	Timeliness (EUCS)	-	−E	-	−E	-	-	−E^{d}	−E	-	−E	-
SK-E-88	EUCS-Skala-1988	-	-	+S^{d}	-	-	-	-	+S	+S^{d}	-	-
SK-E-09	EUCS-Skala-2009	-	-	-	-	+S^{d}	-	-	+S	+S^{d}	-	-
SK-E-13	EUCS-Skala-2013	−E	-	-	-	+S^{d}	-	-	+S	+S^{d}	-	-
SK-G-13	Gesamtskala-2013	-	-	-	+S	+S^{d}	-	−E	+S	+S^{d}	+S	-
SK-K	Kriteriumsskala	-	-	-	-	-	-	-	-	-	-	-
E01	Ich glaube, ich werde in zehn Minuten ... relevante Dokumente finden.	-	-	-	-	-	-	-	-	-	-	+E
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	-	-	-	-	+S^{d}	-	−E	+S −E	+S^{d}	-	+S
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	-	-	-	-	-	-	−E	-	-	-	-
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	−E	−E	+S^{d}	+S	-	-	+S −E	-	+S^{d}	−E	-
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	−S −E	-	-	-	-	-	−S −E	-	-	-	-
E06-M	Erwartungsskala	-	-	+S^{d}	-	+S^{d}	+S^{d}	−E	-	+S^{d}	-	+S

[a] Entspricht auch den Skalen SK15-M und SK19-M.
[b] Entspricht auch der Skala SK18-M.
[c] Entspricht auch der Skala SK13-M.
[d] Es besteht eine Abhängigkeit zwischen Kovariate und abhängiger Variable.
[e] Teilweise ergeben sich negative Gruppenmittelwerte.

Hinzukommen eines Systemeffekts oder einem Hinzukommen eines Interaktionseffekts kommt. Einzig die Variablen E01, E05, F19 und SK15-F, fallen aus diesem Schema heraus, da hier Systemeffekte wegfallen bzw. Erwartungseffekte hinzukommen. Jedoch stimmt die Richtung der hinzukommenden Effekte mit denen der Hauptauswertung überein.

Wie bereits in den einleitenden Absätzen zu Abschnitt 7.4.6.2 dargelegt, sind Benutzerleistungsmaße als Kovariaten besonders anfällig für Abhängigkeiten mit den unabhängigen Variablen, da sie ja gerade die Reaktion der Testteilnehmer auf Systemqualität und Erwartungshaltung

erfassen sollen. In diesem Sinne sind die folgenden Ergebnisse immer unter einem gewissen Vorbehalt zu betrachten, wobei sie natürlich nichtsdestotrotz einen Einblick in den Zusammenhang zwischen Nutzerzufriedenheit und Nutzerleistung ermöglichen. Des Weitern kann davon ausgegangen werden, dass Korrelationen zwischen Kovariate und unabhängigen Variablen lediglich die Wirkung des experimentellen Treatments verringern und somit keine falsch positiven Effekte zu erwarten sind (vgl. Abschn. 4.3.2.3). Um diese Problematik dennoch zu kontrollieren, wird eine mögliche Abhängigkeit wie in den zuvor beschrieben Kovarianzanalysen überprüft. Jedoch werden diese Fälle nicht aus der Betrachtung ausgeschlossen, sondern in den Tabellen 7.30 und 7.31 durch eine Fußnote gekennzeichnet. Im Großen und Ganzen stellt sich diese Abhängigkeit für die Effektivitäts- und Aufwandsmaße am unproblematischsten dar, wohingegen bei Kovariaten, die die Wahrnehmung der Relevanz betreffen, stärker zum Tragen kommt. Dies erscheint nicht weiter verwunderlich, da, um diese Abhängigkeit zu überprüfen, analog zur Hauptanalyse eine Varianzanalyse mit der Kovariaten als abhängiger und Systemgüte und Erwartungshaltung als unabhängigen Variablen durchgeführt wird (vgl. Abschn. 4.3.2.3). In diesem Sinne impliziert ein signifikanter System- oder Erwartungseinfluss im Rahmen der Hauptanalyse also automatisch eine solche Abhängigkeit, wie sie insbesondere für die Relevanzwahrnehmungsmaße B04, B06 und B16 vorliegt, die schon im Rahmen der Hauptanalyse signifikante Effekte zeigen (vgl. Abschn. 7.4.3.1 und 7.4.3.2). Allerdings werden die Resultate für die eine solche Abhängigkeit zu Tage tritt in vielen Fällen durch eine Kovariate die diese Abhängigkeit nicht aufweist, gestützt. Bspw. kommt für die Items F17 und F18 ein zusätzlicher Systemeffekt für die Kovariate B06 hinzu, wobei diese jedoch selbst eine Abhängigkeit von den Treatments zeigt (vgl. Tab 7.30). Für beide Items lassen sich jedoch auch Systemeffekte für die Kovariate B12 nachweisen, die ihrerseits aber unabhängig von den Treatmentgruppen ist.

Zunächst wird nun auf das Hinzukommen bzw. Wegfallen von Haupteffekten eingegangen bevor im Anschluß die neu auftretenden Interaktionseffekte genauer besprochen werden. Wie eingangs erwähnt kommt in der Mehrheit der Fälle ein Systemeffekt hinzu oder es verschwindet ein zuvor signifikanter Erwartungseinfluss. Obwohl dieses Verhalten vereinzelt auch für die Effektivitäts- und Aufwandsmaße zu beobachten ist, treten diese beiden Effekte doch am deutlichsten im Kontext der wahrgenommenen Relevanz zu Tage. Sowohl das Verschwinden des Erwartungseinflusses als auch das neu Auftreten des Systemeffekts lassen sich im Rahmen der bereits besprochenen Anpassungseffekte der Relevanzwahrnehmung verstehen. Für die Erwartungshaltung kann im Rahmen der Hauptanalyse nachgewiesen werden, dass eine höhere Erwartung zu einer weniger restriktiven Relevanzbewertung führt. Den Testteilnehmern mit der höheren Erwartung erscheinen die zurückgegebenen Suchergebnisse also relevanter als den Teilnehmern mit der niedrigeren Erwartungshaltung. Wie in der Hauptanalyse festgestellt geht diese weniger restriktive Relevanzwahrnehmung mit einer höheren Nutzerzufriedenheit einher. Rechnet man diesen Einfluss nun jedoch, wie hier geschehen, aus der Nutzerzufriedenheit heraus so verschwindet der dazugehörige Erwartungseffekt. In diesem Sinne beeinflusst die Selbstwahrnehmung der Eigenleistung bzw. die wahrgenommene Qualität der Suchmaschine also tatsächlich die Zufriedenheit der Nutzer.

Das Hinzukommen des Systemeffekts lässt sich auf eine ähnliche Weise verstehen, wobei hier der im Rahmen dieser Arbeit nachgewiesene systembedingte Anpassungseffekt der Relevanz-

Tab. 7.31.: Übersicht über Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses weiterer leistungsbezogener Störfaktoren auf die Benutzerzufriedenheit in SP_A. Die Tabelle stellt für jede Kovariate alle Zufriedenheitsindikatoren mit entfallenden (–) bzw. hinzukommenden (+) Effekten der Systemleistung (S), der Erwartung (E) oder ihre Interaktion (I) dar.

ID	Beschreibung	M10	M16	M37	M45	S01	S02	S03	S04	S05	S06
F08	Ist die Suchmaschine benutzerfreundlich?	-	-	-	-	-	-	-	-	-	$+I_{SE}$
F12	Ist die Suchmaschine erfolgreich?	$+I_{SE}$	-	$+I_{SE}$	-	-	-	-	-	-	-
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	-	-	-	-	-	-	-	-	-	+S
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	-	-	-	-	-	-	−S	−S	-	−S
F23	Ich bin mit meiner Suchleistung zufrieden.	−E	-	-	−E	-	-	-	-	-	-
SK03-M	Benutzerfreundlichkeit	-	-	-	$+I_{SE}$ +S	-	-	$+I_{SE}$	-	-	-
SK04-M	Suche	-	-	-	+S	-	-	-	-	-	-
SK14-F	Suche	-	-	-	+S	-	-	-	-	+S [a]	-
SK15-F	Eigenleistung	+E	+E	+E	+E	+E	+E	+E	+E	+E	+E
E01	Ich glaube, ich werde in zehn Minuten ... relevante Dokumente finden.	-	-	-	-	+E	-	-	-	+E [a]	-
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	-	−S	-	-	-	-	-	-	-	-
E06-M	Erwartungsskala	-	-	-	+S	-	-	-	-	-	-

[a] Es besteht eine Abhängigkeit zwischen Kovariate und abhängiger Variable.

bewertung als Erklärungsgrundlage dient. Wie im Rahmen der Hauptanalyse festgestellt, führt eine höhere Systemqualität zu einer strengeren Relevanzbewertung der Dokumente, womit die wahrgenommene Systemqualität geringer ausfällt. Im Ergebnis nehmen die Probanden beide System als gleich gut wahr (vgl. Abb 7.17). Dies deckt sich mit dem so gut wie nicht vorhanden Einfluss der Systemleistung im Rahmen der Zufriedenheitsauswertung der Hauptanalyse. Korrigiert man nun jedoch das Zufriedenheitsurteil um die wahrgenommene Systemleistung, so tritt die tatsächliche Systemqualität zu Tage und die Testteilnehmer zeigen sich zufriedener mit dem besseren System. Es kann also erneut nachgewiesen werden, dass die von den Probanden wahrgenommene und durch den Anpassungseffekt korrigierte Systemleistung einen unmittelbaren Einfluss auf das Zufriedenheitsurteil ausübt. In den Fällen hingegen in denen nach der Kovarianzanalyse sowohl ein signifikanter System- also auch ein signifikanter Erwartungseinfluss vorliegt, können die Vorhersagen des C/D-Paradigmas in Bezug auf die positive Konfirmation bestätigt werden. Zwar sind wiederum Nutzer mit der höheren Erwartung generell zufriedener als Nutzer mit der niedrigeren Erwartungshaltung, doch führt darüber hinaus in beiden Gruppen die Nutzung des besseren Systems zu einer höheren Zufriedenheit.

Weiterhin zeigen sich für sieben Zufriedenheitsindikatoren neu auftretende Interaktionseffekte (F08, F12, F25, SK03-M, SK09-M, SK13-M/SK-E u. SK13-F). Zum größten Teil stehen diese Variablen im Zusammenhang mit der Benutzerfreundlichkeit des Testsystems (F08, SK03-M, SK09-M, SK13-M/SK-E u. SK13-F). Dabei ist das Item F08 bereits in allen vier Skalen enthalten. Jedoch ist zu beachten, dass Wechselwirkungseffekte für die vier Zufriedenheitsskalen auch für Benutzerleistungskovariaten auftreten, für die F08 alleine keine signifikante Interaktion zwischen Systemgüte und Erwartungshaltung anzeigt.

Da die Interaktionseffekte für alle Items und Skalen qualitativ ähnlich ausfallen, ist in Abbildung 7.24 exemplarisch das Interaktionsdiagramm für Frageitem F08 dargestellt, welches im Rahmen der Konfidenzintervalle die stärkste Trennung der Treatmentgruppen aufweist. Wie bereits im Rahmen der demographischen und erfahrungsbezogenen Kovariaten beobachtet (vgl.

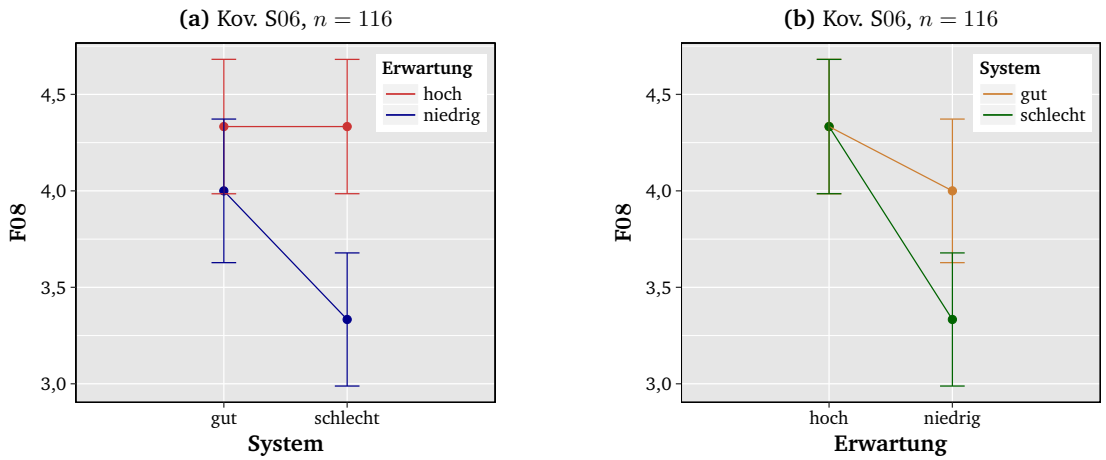


Abb. 7.24.: Wechselwirkung zwischen Systemleistung und Erwartungshaltung für das Frageitem: *Ist die Suchmaschine benutzerfreundlich?* (F08) unter Berücksichtigung der Auffindezeit des ersten richtig als relevant bewerteten Dokuments bei 4-stufiger Relevanzskala S06 als Kovariate. Bild (a) zeigt die Erwartungshaltung in Abhängigkeit der Systemgüte, während Bild (b) die Systemgüte in Abhängigkeit der Erwartungshaltung darstellt. Bild (a): Die hohe Erwartungshaltung führt über beide Systemgüten hinweg tendenziell zu einer höheren Zufriedenheit im Vergleich zur niedrigen Erwartungshaltung. Während sich für die hohe Erwartungshaltung darüber hinaus keine Systemabhängigkeit zeigt, fällt die Zufriedenheit mit dem schlechteren System bei der niedrigen Erwartungshaltung gegenüber dem besseren System ab. Bild (b): Eine Abhängigkeit von der Systemleistung ist für die Zufriedenheit mit der Benutzerfreundlichkeit nur im Kontext der niedrigen Erwartungshaltung erkennbar. Fehlerbalken kennzeichnen 95%-Konfidenzintervalle der Gruppenmittelwerte.

Abb. 7.23), dreht sich der allgemeine Trend, dass höhere Erwartungen zu einer größeren Zufriedenheit bei den Nutzern führen nicht um (Bild (a)). In diesem Sinn kann das C/D-Paradigma also weiterhin nicht bestätigt werden. Allerdings ergibt sich durch die Interaktion zwischen Systemgüte und Erwartungshaltung nun erneut ein differenzierteres Bild. So ist es für alle sieben Zufriedenheitsindikatoren der Fall, dass innerhalb der niedrigeren Erwartungshaltung Nutzer des besseren Systems weitaus zufriedener als Nutzer des schlechteren Systems sind (Bild (b)). Dieser Aspekt der positiven Diskonfirmation des C/D-Paradigmas ist also wiederum zu beobachten. Anders stellt sich die Situation hingegen im Kontext der hohen Erwartungshaltung dar, wo kein Unterschied zwischen den beiden Systemgüten zu erkennen ist. Im Fall der hohen Erwartungshaltung scheint der Systemunterschied von den Probanden also nicht wahrgenommen zu werden.

Eine mit den Ergebnissen der Benutzerleistungsauswertung konsistente Erklärung dieses Verhaltens ergibt sich wiederum unter Berücksichtigung der in Abschnitt 7.4.3.2 dargestellten Interaktionseffekte in Bezug auf die 4-stufige Relevanzskala (M27, V38 u. V57). Dort ist zu beobachten, dass der systemgütenbasierte Anpassungseffekt im Wesentlichen nur im Kontext der hohen Erwartungshaltung zu beobachten ist: Nutzer mit der hohen Erwartungshaltung wenden mit steigender Systemleistung strengere Relevanzkriterien an. Im Fall der niedrigen Erwartungshaltung hingegen lässt sich keine Adaption der Relevanzwahrnehmung feststellen. Diese erwartungsspezifische Wahrnehmung der Systemleistung spiegelt sich nun auch in der Zufriedenheitsantwort der Probanden wider. Ohne den systembedingten Anpassungseffekt, d.h. im Fall einer niedrigen

Erwartungshaltung, ist positive Diskonfirmation zu beobachten. Nutzer des besseren Systems sind also zufriedener als Nutzer des schlechteren Systems. Kommt hingegen der systembedingte Anpassungseffekt zum Tragen, was im Fall der hohen Erwartungshaltung geschieht, ist hingegen kein Einfluss der Systemgüte erkennbar. Dies kann wiederum als Folge der wahrgenommenen Qualität der Suchmaschine interpretiert werden, da durch den Anpassungseffekt der Relevanzwahrnehmung die Qualität der beiden Suchmaschinen von den Probanden als identisch wahrgenommen wird. Die Berücksichtigung einiger der hier gewählten Benutzerleistungskovariaten scheint also dazu zu führen, dass die im Rahmen der Auswertung der Benutzerleistung nachgewiesenen Anpassungseffekte nun auch im Zufriedenheitsurteil der Probanden sichtbar werden. Insbesondere tritt dieser Effekt erneut für Kovariaten auf, die im Zusammenhang mit der Wahrnehmung der eigenen Suchleistung stehen (M10, B17 u. B18). In diesen Fällen scheint also das Herauspartialisieren der wahrgenommenen Eigenleistung ausschließlich in Bezug auf die niedrige Erwartungshaltung zu gelingen, so dass dort der Systemeinfluss sichtbar wird, während der Einfluss des systembedingten Anpassungseffekts bei der hohen Erwartung nach wie vor den Systemunterschied unterdrückt.

Zusammenfassend zeigt sich somit, dass die Einbeziehung leistungsbezogener Kovariaten in die Zufriedenheitsauswertung interessante Informationen liefert. Zwar kann das C/D-Paradigma weiterhin nicht in vollem Umfang bestätigt werden, allerdings zeigt sich, wie im Fall der demographischen und erfahrungsbezogenen Kovariaten eine positive Diskonfirmation im Kontext der niedrigen Erwartungshaltung. Darüber hinaus ergeben sich durch diese spezielle Art der Kovarianzanalyse Hinweise auf den Einfluss des systembedingten Anpassungseffekts auf die Zufriedenheit der Testpersonen: Wird die wahrgenommene Relevanz der Suchergebnisse, die mit steigender Systemleistung geringer ausfällt, herauspartialisiert, so kann für viele Zufriedenheitsitems und Skalen nun ein positiver Zusammenhang zwischen Systemleistung und Benutzerzufriedenheit nachgewiesen werden.

7.5. Fazit: Experiment 3

Die dritte im Rahmen der vorliegenden Arbeit durchgeführte Nutzerstudie kann als Synthese der beiden vorangegangenen Experimente verstanden werden, wobei gleichzeitig der Fokus der Untersuchung nicht mehr auf den globalen Systemgüten- und Erwartungseinfluss gerichtet ist, sondern sich in Richtung der zeitlichen Entwicklung des Nutzerverhaltens verschiebt. Dabei werden auf der einen Seite erfolgreiche Testdesignelemente, wie bspw. die freie Interaktion mit dem Suchsystem, die verwendeten Erhebungsinstrumente sowie das Auswertungskonzept aus Experiment 2 übernommen, auf der anderen Seite jedoch die Komplexität des Untersuchungsdesigns reduziert, indem den Testpersonen analog zu Experiment 1 lediglich eine Systemgüte präsentiert wird. Die Abfrage der dynamischen Komponente der Nutzerzufriedenheit wiederum führt zu einem erhöhten Aufwand für die Probanden, da nun das Zufriedenheitsurteil nach jeder bearbeiteten Suchaufgabe erfragt werden muss. In Bezug auf die Manipulation der Erwartungshaltung wird der erfolgreiche Ansatz aus Experiment 2 übernommen, im Kontext der Testinstruktion auf einen Systemvergleich hinzuweisen und somit die zu erwartende Systemleistung zu motivieren, wobei dieses Mal auf eine audiovisuelle Darbietung des Stimulusmaterials zurückgegriffen wird. Darüber hinaus erlaubt die Untersuchung im Zuge der erweiterten, 4-stufigen Relevanz-

skala neben einer Betrachtung der dynamischen Entwicklung des Nutzerverhaltens auch einen differenzierteren Blick auf die Relevanzwahrnehmung der Probanden.

Im Kontext der Benutzerleistung können im Rahmen der Mittelwertsauswertung alle Befunde der beiden vorangegangenen Nutzerstudien bestätigt werden. Dies betrifft sowohl den Kompensationseffekt in Bezug auf die Recallmaße, wie auch die Adaption der Relevanzwahrnehmung, verursacht durch Systemgüte und Erwartungshaltung. Nach wie vor ist auf der einen Seite mit steigender Systemqualität die Verwendung restriktiverer Relevanzkriterien durch die Testteilnehmer zu beobachten, wohingegen auf der anderen Seite eine höhere Erwartungshaltung mit einer weniger restriktiven Relevanzbewertung verknüpft zu sein scheint. Dabei ist letzterer Effekt jedoch größtenteils nur im Rahmen der besser kontrollierten Stichprobe SP_B sichtbar, wohingegen die systemleistungsbedingten Effekte auch für die weniger kontrollierte Gesamtstichprobe nachweisbar sind. Interessanterweise zeigt sich darüber hinaus, dass die restriktiveren Relevanzkriterien bei höherer Systemleistung mit erhöhten Betrachtungszeiten und somit einer verminderten Sucheffizienz der Probanden einherzugehen scheint. Der Mehrwert der in Experiment 3 neu verwendeten 4-stufigen Relevanzskala wird an dem detaillierteren Einblick in die Wirkungsweise dieser Effekte deutlich. So zeigt sich zum einen, dass die beiden Anpassungsreaktionen im Wesentlichen in Bezug auf die beiden mittleren Relevanzkategorien *eher relevant* und *eher irrelevant* zum Tragen kommt, während sich die beiden extremen Relevanzstufen *relevant* und *irrelevant* als weitestgehend unabhängig von Systemgüte und Erwartungshaltung herausstellen. Des Weiteren werden im Kontext der 4-stufigen Relevanzskala nun auch Wechselwirkungen zwischen Systemleistung und Erwartungshaltung signifikant, die darauf hindeuten, dass das systembedingte Anpassungsverhalten im Wesentlichen auf die Probanden mit hoher Erwartungshaltung beschränkt bleibt. In Bezug auf das dynamische Nutzerverhalten ist zunächst festzustellen, dass alle bisher beschriebenen Effekte unter Berücksichtigung der wiederholten Interaktion der Probanden mit dem Testsystem weiterhin Bestand haben. Darüber hinaus lassen sich für einzelne Benutzerleistungsmaße Lern- und Ermüdungseffekte nachweisen. Zeitliche Abhängigkeiten der beiden Faktoren Systemgüte und Erwartungshaltung treten hingegen ausschließlich für eine Auswertung mit klassischen varianzanalytischen Verfahren auf, bei denen die statistischen Voraussetzungen nicht durchgehend erfüllt sind. Trotzdem lässt sich auf diese Weise ein erster Eindruck der Dynamik dieser Effekte gewinnen. So deuten die Resultate zum einen darauf hin, dass sich der systembedingte Anpassungseffekt erst nach einer gewissen Arbeitszeit mit dem System einstellt, sich dann jedoch über die Zeit noch weiter zu verstärken scheint. Darüber hinaus finden sich Hinweise auf dynamische Effekte, die innerhalb der einzelnen Suchsessions zum Tragen kommen.

Wie schon für die Benutzerleistung bestätigt die Mittelwertanalyse auch für die Benutzerzufriedenheit die Hauptergebnisse des zweiten Experiments. So kann insbesondere die Erwartungsmanipulation erneut als erfolgreich angesehen werden, da die überwiegende Mehrheit der betrachteten Zufriedenheitsindikatoren eine signifikante Abhängigkeit von der Erwartungshaltung zeigt. Des Weiteren setzt sich der allgemeine Trend fort, dass im Gegensatz zu den Vorhersagen des C/D-Paradigmas eine höhere Erwartungshaltung zu einem positiveren Zufriedenheitsurteil führt. Demgegenüber fällt der Einfluss der Systemgüte weitaus geringer aus, wobei jedoch weiterhin die Benutzerzufriedenheit mit steigender Systemleistung zunimmt. Auch in Be-

zug auf die Dynamik ist für die Benutzerzufriedenheit eine schwächere Abhängigkeit als für die Benutzerleistung feststellbar. So ist nur für ein einzelnes Frageitem eine signifikante Abhängigkeit des Zufriedenheitsurteils vom Bearbeitungszeitpunkt nachweisbar. Hinweise darauf, dass die Abwesenheit eines Einflusses der Systemgüte ursächlich mit der Anpassung der Relevanzkriterien zusammenhängt, die zu einer Verringerung der wahrgenommenen Systemunterschiede beitragen, finden sich im Rahmen der Kovarianzanalyse. Wird die Relevanzwahrnehmung der Testteilnehmer aus den Zufriedenheitsreaktionen herauspartialisiert, ergibt sich in vielen Fällen ein signifikanter Einfluss der Systemgüte, bei der die Systemqualität positiv mit der Benutzerzufriedenheit korreliert. Im Rahmen der signifikanten Wechselwirkungen zwischen Systemleistung und Erwartungshaltung wird darüber hinaus deutlich, dass sich ein signifikanter Einfluss der Systemleistung nur im Kontext der niedrigen Erwartungshaltung einstellt, bei der der die wahrgenommene Systemleistung verringernde Anpassungseffekt der Relevanzkriterien im Gegensatz zur hohen Erwartungshaltung gerade nicht zum Tragen kommt.

Zusammenfassend lässt sich somit festhalten, dass die Untersuchungsziele der dritten Nutzerstudie erreicht werden. Zum einen können die Befunde der beiden vorangegangenen Experimente erneut bestätigt werden. Zum anderen erlaubt das erweiterte Studiendesign anhand der 4-stufigen Relevanzskala und der dynamischen Erfassung des Nutzerverhaltens detaillierte Erkenntnisse in Bezug auf die Wirkungsmechanismen und Abhängigkeiten dieser Effekte, die wiederum als Grundlage weitergehender Untersuchungen dienen können.

8. Zusammenfassung und Interpretation der zentralen Ergebnisse

Ziel dieses Kapitels ist es, die auf Basis der Einzelstudien gewonnenen Erkenntnisse aus einer vergleichenden Perspektive zu betrachten und dabei Differenzen und Überschneidungen zwischen ihnen herauszuarbeiten. Den Ausgangspunkt bilden dabei die drei zentralen Forschungsfragen, die im Rahmen dieser Arbeit untersucht werden: (1) Lassen sich die Implikationen des C/D-Paradigmas auf den IR-Kontext übertragen? (2) Führt eine höhere Systemleistung zu höherer Zufriedenheit und Benutzerleistung? (3) Wie ändert sich die Wahrnehmung der Suchmaschinenqualität dynamisch im Suchprozess? Dabei erlaubt das gewählte Forschungsdesign die unabhängigen Größen Systemgüte und Erwartungshaltung unter kontrollierten Bedingungen zu variieren, um ihre Wirkung auf das Nutzerverhalten zu analysieren. Ausgehend von den im Stand der Forschung dargestellten Grundlagen zur Wahrnehmung von Suchergebnissen (vgl. Kap. 2 u. 3), in denen Nutzererwartungen als zentrale Einflussgröße identifiziert werden, kommt der Untersuchung der Erwartungshaltung ein besonderes Interesse zu. Aus informationswissenschaftlicher Perspektive interessieren insbesondere auch die verwendeten Methoden und deren jeweilige Rahmenbedingungen, mit deren Hilfe sich die gestellten Forschungsfragen beantworten lassen.

Aufbauend auf den zentralen Befunden der Kapitel 5 bis 7 werden im Folgenden die experimentellen Ergebnisse der Nutzerstudien noch einmal explizit in Bezug auf die forschungsleitenden Fragen diskutiert. Zur besseren Einordnung der jeweiligen Ergebnisse wird darüber hinaus eine Einteilung der Befundlage in über die Gesamtdaten gesicherte Erkenntnisse und im Zuge der konkreten Untersuchungen gewonnene Einzelerkenntnisse vorgenommen. Ähnlich wie im Fall des in den letzten beiden Experimenten zum Einsatz kommenden stichprobendifferenzierenden Auswertungskonzeptes wird damit auch hier von einem fortlaufenden Prozess der Erkenntnisgewinnung ausgegangen, in dessen Verlauf die empirische Belastbarkeit der Befunde im Sinne einer Kreuzvalidierung gestützt wird. Diese Unterscheidung macht jedoch auch deutlich, in welchen Bereichen weiterer Forschungsbedarf besteht und welche Anforderungen sich dabei an die Forschungspraxis ergeben.

8.1. Forschungsfrage FF1: Übertragbarkeit des C/D-Paradigmas

Die erste Forschungsfrage adressiert die Auswirkungen von Benutzererwartungen auf die Qualitätswahrnehmung von Suchergebnissen. Leitend ist in diesem Zusammenhang die Frage nach der Übertragbarkeit der aus der Kundenzufriedenheitsforschung stammenden Diskonfirmationstheorie auf den Kontext der Informationssuche. Da der Einfluss von Nutzererwartungen auf den Suchprozess in der IR-Forschung bisher wenig Beachtung findet, wird im Zuge der Experimentplanung ein iteratives Vorgehen angewandt, bei dem der Erkenntnisgewinn der einzelnen Experimente zur sukzessiven Weiterentwicklung der Testmaterialien der nachfolgenden Experi-

mente genutzt wird. Bezüglich der für die erste Forschungsfrage zentralen Einflussnahme der Erwartungshaltung können somit drei unterschiedliche Manipulationsansätze verglichen werden. Im Folgenden werden die Ergebnisse aller drei Experimente aufeinander bezogen.

8.1.1. Über die Gesamtdaten gesicherte Erkenntnisse zu FF1

Einen der theoretischen Ausgangspunkte der vorliegenden Arbeit bildet die Frage, ob sich die formale Ähnlichkeit zwischen dem Prozess der Informationssuche und der Inanspruchnahme einer Dienstleistung auch im Kontext der Zufriedenheitsbildung und insbesondere in Bezug auf einen ähnlich gearteten Erwartungs-Wahrnehmungs-Vergleich widerspiegelt. Vor dem Hintergrund der in den Kapiteln 5 bis 7 dargestellten Ergebnisse kann diese These jedoch in keiner der drei Untersuchungen bestätigt werden. So lässt sich zwar im Rahmen des zweiten und dritten Experiments ein signifikanter Einfluss der Erwartungshaltung auf viele der getesteten Zufriedenheitsindikatoren nachweisen, jedoch zeigt sich zumindest für die Erwartungshaltung nicht das durch das C/D-Paradigma prognostizierte Verhalten. Stattdessen wird deutlich, dass der Zusammenhang zwischen Benutzererwartungen und - zufriedenheit komplexer ist als zunächst angenommen. So lassen die signifikanten Ergebnisse hinsichtlich der Zufriedenheit der Probanden im zweiten und dritten Experiment zunächst den Schluss zu, dass die Testteilnehmer in ihren Zufriedenheitsurteilen im Wesentlichen ihre Erwartungsmanipulation widerspiegeln, ohne, dass ein starker Einfluss der Systemleistung erkennbar wäre. Ein mögliches Erklärungsmodell liefert das Auftreten eines Placebo-Effekts bzw. Bestätigungsfehlers (confirmation bias), bei dem die Probanden bestrebt sind, ihre Erwartungen zu bestätigen und gegenteilige Wahrnehmungen auszublenden. Auf ein solches Verhalten wird bspw. auch in einer Studie von Habel et al. (2016) aus dem Bereich der Kundenzufriedenheitsforschung hingewiesen.

Diese Interpretation lässt jedoch den beobachteten Einfluss der Erwartungshaltung auf die Relevanzwahrnehmung der Testpersonen außer Acht, weshalb in der vorliegenden Arbeit ein zweiter Erklärungsansatz bevorzugt wird, der über den klassischen Soll-Ist-Vergleich von erwarteter und wahrgenommener Leistung hinaus auch den Entstehungsprozess der zu bewertenden Leistung berücksichtigt. Tatsächlich deuten die durchgeführten Analysen darauf hin, dass die Ursprungsvermutung, die aktive Mitarbeit des Nutzers könnte einen maßgeblichen Einfluss auf die Zufriedenheitsbeurteilung haben, von den vorliegenden Untersuchungsdaten gestützt wird. In Experiment 2 kann festgestellt werden, dass die Wahrnehmung der eigenen Suchleistung bei hohen Erwartungen ebenfalls höher ausfällt, was in der Folge auch zu höheren Zufriedenheitsurteilen für das Testsystem geführt haben könnte. Diese Schlussfolgerung wird durch die in Experiment 3 gewonnen Ergebnisse untermauert, die insbesondere zeigen, wie Systemqualität, Relevanzwahrnehmung und Zufriedenheitsbeurteilung aufeinander bezogen sind. Konkret kann ein durch die Erwartungshaltung gesteuerter Anpassungseffekt der Relevanzwahrnehmung nachgewiesen werden, der bei hoher Erwartungsmanipulation zu einer weniger restriktiven Relevanzbeurteilung führt. Es erscheint somit nicht verwunderlich, dass in diesem Fall auch das gesamte System von den Teilnehmern positiver wahrgenommen und beurteilt wird. Interessanterweise lässt sich diese positivere Sichtweise der Probanden auf das Testsystem auch für relevanzwahrnehmungsunabhängige Aspekte, wie die Benutzerfreundlichkeit, nachweisen. Auf die Frage, ob die eingesetzte Erwartungsmanipulation letztlich zu einer Änderung der normativen oder prädiktiven Erwartungen der Testpersonen führt, wird genauer in Abschnitt 8.2.2.2 im Kontext des Einflusses der

Anpassungseffekte der Relevanzwahrnehmung auf die Benutzerzufriedenheit eingegangen.

Zusammenfassend lässt sich aus den Gesamtergebnissen der drei Untersuchungen somit schließen, dass sich aufgrund der hier durchgeführten Nutzerstudien für den IR-Kontext hinsichtlich der Erwartungshaltung keine der von dem C/D-Paradigma vorhergesagten diskonfirmatorischen Effekte nachweisen lassen. Vielmehr deutet die positive Korrelation zwischen Zufriedenheitsurteil und Erwartungshaltung zusammen mit dem beobachteten erwartungsbedingten Anpassungsverhalten in Bezug auf die Relevanzwahrnehmung auf die Bedeutung der Wahrnehmung der eigenen Suchleistung für die Zufriedenheitsbildung im Rahmen der Informationssuche hin. Der folgende Abschnitt geht noch einmal dediziert auf einzelne Aspekte der Erwartungsmanipulation im Hinblick auf die drei Nutzerstudien ein.

8.1.2. Im Zuge der Untersuchungen gewonnene Einzelerkenntnisse zu FF1

Eine wesentliche Voraussetzung für die Untersuchung der Übertragbarkeit des C/D-Paradigmas auf den Kontext der Informationssuche ist die Manipulation der Erwartungshaltung. Laborexperimente zeichnen sich dadurch aus, dass die unabhängigen Variablen zumindest prinzipiell beliebig manipuliert werden können. Allerdings erlaubt dies zunächst noch keine Aussage über die resultierende Effektstärke für die abhängigen Variablen. Es gilt daher, eine Manipulationsstrategie zu finden, die einerseits stark genug ausfällt, um den Einfluss der Erwartungshaltung sichtbar zu machen, und die anderseits unauffällig genug ist, um von den Probanden nicht bewusst wahrgenommen zu werden. Im Rahmen dieser Arbeit erfolgt die Manipulation der Erwartungshaltung über Instruktionstexte, die den Probanden zu Beginn der Experimente auf unterschiedliche Weise dargeboten werden. Die Auswertung des ersten Experiments ergibt, dass die Manipulation der Erwartungshaltung der Testpersonen keinen Einfluss auf die Wahrnehmung der Suchergebnisse hat. Es wird daher vermutet, dass in diesem Fall die Stärke der experimentellen Manipulation nicht ausreicht, um eine messbare Veränderung der abhängigen Variablen zu erreichen.

Die im Zuge der Folgeexperimente ergriffenen Maßnahmen, um die Wirkung der Erwartungsmanipulation zu erhöhen und mögliche Verzerrungen durch die Darbietung der Instruktionen zu minimieren, zeigen jedoch, dass der eingeschlagene Kurs der Verwendung von Instruktionstexten nicht grundsätzlich in Frage zu stellen ist. In der Tat führt in den beiden späteren Experimenten die Erwartungsmanipulation zu einer signifikanten Änderung des Zufriedenheitsurteils. Dabei tragen hohe Erwartungen zu einer signifikant höheren Zufriedenheitsreaktion bei, wodurch die Hypothese, dass Erwartungen einen direkten Einfluss auf die Benutzerzufriedenheit zeigen, bestätigt wird. Insbesondere die gewählte Strategie der vergleichenden Evaluierung mehrerer Systeme im Kontext simulierter Arbeitsaufgaben ist für die Forschungspraxis der Informationswissenschaft hervorzuheben, da sie den Probanden die Einnahme der jeweiligen Erwartungshaltung zu erleichtern scheint und demzufolge höchstwahrscheinlich die Wirksamkeit der experimentellen Manipulation gesteigert wird. Dabei ist die Einfachheit der gewählten Manipulationsstrategie ebenso wichtig wie eine ausreichende Manipulationsstärke. Damit sie von den Probanden möglichst einheitlich aufgenommen wird, sollte die verwendete Hintergrundgeschichte deshalb nicht zu lang, jedoch inhaltlich schlüssig und einfach zu erinnern sein. Die ambivalenten Ergebnisse von Experiment 2 deuten in diesem Zusammenhang darauf hin, dass eine Variation der Erwartungshaltung innerhalb der Testpersonen möglicherweise zu komplex ist, um die Wirkung der Manipulation über die Dauer des gesamten Experiments aufrecht zu erhalten. Demgegenüber

zeigen die Ergebnisse von Experiment 3, dass es zu genügen scheint, den Systemvergleich im Rahmen der Instruktion zu erwähnen, um bei den Probanden die intendierte Erwartungshaltung hervorzurufen. Tatsächlich stellt der angekündigte Systemvergleich zwischen zwei Suchsystemen, ausgehend von demselben 2×2-Versuchsplan, ein wesentliches Unterscheidungskriterium zu der im ersten Experiment verfolgten, nicht erfolgreichen Manipulationsstrategie dar.

Darüber hinaus zeigt sich, dass neben der konkreten Manipulationsstrategie auch die Darbietung der Instruktion zum Erfolg der Manipulation beiträgt. Diesbezüglich ist es besonders wichtig, die Aufmerksamkeit der Probanden zu erreichen, um sicherzustellen, dass das Treatment auch wirklich zur Kenntnis genommen wird. Hier haben die Erfahrungen aus dem ersten Experiment dazu geführt, dass im zweiten und dritten Experiment anstelle einer schriftlichen Instruktion eine audiovisuelle Form der Darbietung gewählt wird, bei der die Aufmerksamkeit der Testpersonen durch die gleichzeitige Ansprache des Hör- und Sehsinns stärker in Anspruch genommen wird. Der signifikante Einfluss der erhaltenen Erwartungsmanipulation in beiden Experimenten legt nahe, dass auch diese zusätzlich ergriffenen Maßnahmen zum Erfolg der in dieser Arbeit entwickelten Manipulationsmethode beigetragen haben.

Vor dem Hintergrund und in der Replik dieser Einzelergebnisse wird deutlich, dass das Gelingen einer wirksamen Erwartungsmanipulation im Kontext experimenteller IR-Studien am besten durch einen Ansatz einer den Vergleich mehrerer Systeme in den Blick nehmenden Manipulationsstrategie – unterstützt durch eine audiovisuelle Darbietung der Testinstruktion – ermöglicht werden kann.

8.2. Forschungsfrage FF2: Einfluss der Systemqualität

Wie bereits an verschiedenen Stellen diskutiert, stellt sich mit Blick auf die lange Tradition systemorientierter Forschung im Kontext der IR-Evaluierung die Frage nach der Übertragbarkeit systemorientierter Ergebnisse auf den Anwendungsfall. Diesem Sachverhalt wird in der zweiten Forschungsfrage nachgegangen, indem der Einfluss unterschiedlicher Systemgüten auf Benutzerleistung und Zufriedenheit untersucht wird. Darüber hinaus ermöglicht es das im Rahmen dieser Arbeit entwickelte Untersuchungsdesign, zu überprüfen, ob dieser Einfluss von der Erwartungshaltung der Testpersonen moderiert wird. Im dritten Experiment schließlich richtet sich der Fokus dieser Forschungsfrage auf die in den ersten beiden Experimenten beobachtete Anpassung der Relevanzkriterien an die vorgefundene Systemqualität. Eine zentrale Rolle spielt dabei die Einführung einer feiner abgestuften Relevanzskala, die einen detaillierteren Einblick in die dem Anpassungseffekt zugrunde liegenden Mechanismen ermöglicht. Im folgenden Abschnitt werden die wichtigsten Aspekte dieser Ergebnisse noch einmal zusammengefasst.

8.2.1. Über die Gesamtdaten gesicherte Erkenntnisse zu FF2

Um die reale Suchsituation bei der Durchführung experimenteller Studien zum Informationsverhalten näherungsweise abbilden zu können, muss eine Vielzahl von Einflussfaktoren berücksichtigt werden, die zum Teil noch nicht genau bekannt und erforscht sind. Als Forscher bewegt man sich deshalb stets im Spannungsfeld zwischen Experiment und Anwendungsfall. Je mehr die Rahmenbedingungen eines Experiments kontrolliert werden, desto größer ist die Gefahr, sich von der realen Anwendungssituation zu entfernen. Umgekehrt sind die spezifischen Effekte, denen das

Forschungsinteresse gilt, in einem zu offenen Versuchsaufbau u.U. nicht eindeutig nachweisbar. Durch die erneute Überprüfung der im Rahmen eines kontrollierten Laborexperiments gefundenen Ergebnisse an einer anderen Stichprobe kann die Reliabilität der beobachteten Effekte jedoch validiert werden. Im Folgenden wird aus diesem Grund, wie auch schon im Kontext der Ergebnisse zu FF1, zwischen den Ergebnissen, die über alle drei Untersuchungen Bestand haben, und den im Zuge der konkreten Untersuchungen gewonnenen Einzelerkenntnissen, die zu diesem Zeitpunkt nur einen begrenzt generalisierbaren Charakter haben, unterschieden.

8.2.1.1. Beeinflussung der Benutzerleistung durch die Systemqualität

In diesem Abschnitt werden Effekte berichtet, die vor dem Hintergrund der in den Kapiteln 5 bis 7 dargestellten Ergebnisse, auf eine Beeinflussung der Benutzerleistung durch die Systemqualität hindeuten. Ganz allgemein lässt sich feststellen, dass sich die Systemqualität in der Regel positiv auf die Benutzerleistung auswirkt. Darüber hinaus können hinsichtlich der Benutzerleistung zwei Effekte als gesichert gelten. Hierbei handelt es sich zum einen um einen Kompensationseffekt für recallorientierte Benutzerleistungsmaße sowie zum anderen um einen Anpassungseffekt der Relevanzwahrnehmung, der sich in einem systemleistungsbedingten Unterschied in Bezug auf einige der precisionorientierten Benutzerleistungsmaße äußert. Beide Effekte werden im Folgenden genauer diskutiert.

Die kurzfristige Befriedigung von Informationsbedürfnissen ist prinzipiell auch ohne die Unterstützung durch eine ausgefeilte Suchmaschine mit hoher Systemleistung möglich, da Benutzer mit entsprechendem Aufwand auch bei geringerer Systemqualität in der Lage sind, eine vergleichbare Suchleistung zu erreichen. In allen drei Experimenten äußert sich dieses kompensatorische Suchverhalten der Benutzer in den die Vollständigkeit der Suchergebnisse bewertenden recallorientierten Benutzerleistungsmaßen, für die keine signifikanten Unterschiede zwischen den betrachteten Systemleistungsunterschieden nachweisbar sind. Diese Ergebnisse stehen im Einklang mit anderen Studien (vgl. Abschn. 3.2.1, Al-Maskari et al. 2008b; Turpin und Hersh 2001; Allan et al. 2005), bei denen für vergleichbare relative Systemunterschiede ein analoger Kompensationseffekt beobachtet wird. Demgegenüber finden Turpin und Scholer (2006), interessanterweise in einer vergleichbaren Studie für denselben Systemunterschied wie in dieser Arbeit, einen signifikanten Einfluss der Systemgüte auf den Benutzerrecall. In diesem Fall beträgt die Bearbeitungszeit jedoch lediglich fünf Minuten pro Aufgabe, was darauf hindeuten könnte, dass diese relativ kurze Zeitspanne nicht ausgereicht haben könnte, um den vorhandenen Systemunterschied auszugleichen. Es wäre daher interessant, diese Diskrepanz in einer Folgestudie, bei der neben der Systemleistung auch die Bearbeitungszeit für die einzelnen Aufgaben variiert wird, eingehender zu untersuchen. Im Sinne einer Weiternutzung der vorliegenden Untersuchungsdaten bestünde alternativ auch die Möglichkeit, die Bearbeitungszeit im Nachhinein auf fünf Minuten zu beschränken, worauf an dieser Stelle jedoch aus Gründen des Umfangs der Darstellung verzichtet wird.

Der zweite Effekt betrifft die precisionorientierten Benutzerleistungsmaße. Wie erwartet, zeigt sich hier für viele der betrachteten Leistungsmaße, dass eine höhere Systemleistung zu einer höheren Benutzerleistung führt. In Bezug auf diesen Aspekt der Informationssuche profitieren Nutzer also in der Tat von einer höheren Leistung des verwendeten Suchsystems. Zugleich finden sich aber auch einige Precisionmaße, die dieser These zunächst zu widersprechen scheinen, da

eine höhere Systemleistung in diesen Fällen zu einer signifikant niedrigeren Benutzerleistung führt. In diese Kategorie fallen insbesondere Precision- und Imprecisionmaße, die die Übereinstimmung zwischen Juroren- und Nutzerurteil anhand der Relevanzbewertung der Dokumente quantifizieren. Dabei zeigt sich, dass Nutzer des schlechteren Systems eher bereit sind, irrelevante Dokumente als relevant zu akzeptieren, während eine höhere Systemleistung dazu führt, dass die Teilnehmer mehr relevante Dokumente als irrelevant verwerfen. Dies kann als eine Anpassung der angewendeten Relevanzkriterien an die präsentierte Suchmaschinenqualität gewertet werden, bei der eine höhere Systemleistung mit einer restriktiveren Relevanzbewertung einhergeht. Dieser Effekt kann stabil über alle Untersuchungen und Stichprobenqualitäten sowie hinsichtlich einer ganzen Reihe verschiedener Benutzerleistungsmaße nachgewiesen werden. Vor dem Hintergrund, dass ein ähnliches Verhalten bspw. auch in den Studien von Wang und Soergel (1998) und Smucker und Jethani (2010a) beschrieben ist, kann dieser Effekt als gesichert gelten. Die im dritten Experiment beobachteten signifikanten Unterschiede in Bezug auf die Bewertungszeiten fügen diesem Anpassungseffekt eine weitere Dimension hinzu. Es zeigt sich, dass Benutzer des besseren Systems zur Anwendung der strengeren Relevanzkriterien zusätzliche Zeit benötigen, bevor sie Dokumente im Widerspruch zu den Juroren bewerten bzw. allgemein als irrelevant verwerfen. Die Anwendung strengerer Relevanzkriterien scheint also mit zusätzlichem Aufwand für die Nutzer verbunden und erfordert längere Lesezeiten pro Dokument.

In der Gesamtschau lassen sich somit hinsichtlich der Benutzerleistung zwei Hauptaussagen festhalten, die vor dem Hintergrund ihrer Stabilität über alle drei Untersuchungen hinweg als gesichert gelten können und sich darüber hinaus in den Kontext der aktuellen Forschungsliteratur einfügen. Dies betrifft zum einen die beschriebene Kompensationsfähigkeit von Suchmaschinennutzern für recallorientierte Benutzerleistungsmaße und zum anderen die Anwendung restriktiverer Bewertungsstrategien mit steigender Systemleistung.

8.2.1.2. Beeinflussung der Benutzerzufriedenheit durch die Systemqualität

Nachdem der vorangegangene Abschnitt Ergebnisse diskutiert, die auf eine Beeinflussung der Benutzerleistung durch die Systemqualität hindeuten, werden im Folgenden generalisierbare Erkenntnisse in Bezug auf die Benutzerzufriedenheit abgeleitet.

Ein erstes Resultat betrifft die Eignung des EUCS-Instruments (Doll et al., 1994; Doll u. Xia, 1997) für den IR-Kontext, das für die Nutzerstudien dieser Arbeit ins Deutsche übertragen wird (vgl. Abschn. 4.2.2.3). Zusammenfassend lässt sich sagen, dass die theoretische Faktorenstruktur des EUCS-Instruments in den im zweiten und dritten Experiment durchgeführten explorativen Faktorenanalysen nur in Teilen repliziert werden kann. Die inhaltliche Analyse der einzelnen Items lässt jedoch erkennen, dass das Instrument insbesondere bei der Systemleistung, etwa zur Bewertung von Recall und Precision, aber auch hinsichtlich der Wahrnehmung des eigenen Sucherfolgs noch Verbesserungspotential besitzt. Durch zusätzlich entwickelte Items können im Rahmen dieser Arbeit Anhaltspunkte für weitere Dimensionen der Benutzerzufriedenheit gewonnen und einer ersten Prüfung unterzogen werden. Die auf dieser Basis faktoranalytisch ermittelten Skalen lassen sich inhaltlich gut interpretieren und sind ausreichend reliabel.

Über dieses methodische Ergebnis hinaus, kann zumindest vereinzelt ein positiver Zusammenhang zwischen Systemleistung und Nutzerzufriedenheit, insbesondere für precisionorientierte Frageitems, nachgewiesen werden. In diesem Sinne bestätigen sich bspw. die Ergebnisse von

Johnson et al. (2003) und Kelly et al. (2007), dass Nutzer in der Lage sind, Systemunterschiede wahrzunehmen und die Wahrnehmung über ihr Zufriedenheitsurteil zu artikulieren (vgl. Abschn. 3.3.2.2). Für die meisten Zufriedenheitsitems zeigt sich jedoch über alle drei Studien hinweg keine signifikante Abhängigkeit von der Systemleistung. Die Zufriedenheit mit dem schlechteren System ist also genauso hoch wie bei der besseren Systemleistung. Vor dem Hintergrund des Kompensationseffekts in Bezug auf den Benutzerrecall könnte es sich dabei um einen durch die von Dostert und Kelly (2009) und Al-Maskari et al. (2006) beschriebene Informationssättigung hervorgerufenen Effekt handeln, bei dem die Teilnehmer unabhängig von der Systemgüte eine ausreichende Anzahl an relevanten Dokumenten finden, um subjektiv zu der Einschätzung zu gelangen, die Suchaufgabe erfolgreich bearbeitet zu haben. Im Ergebnis stellt sich somit unabhängig von der Systemleistung Zufriedenheit ein. Dabei bleibt jedoch die Frage offen, ob es sich bei dem beschriebenen Sättigungseffekt um ein Artefakt experimenteller Untersuchungsdesigns durch extern vorgegebene Suchaufgaben handeln könnte.

Ein ergänzender Erklärungsansatz für die Abwesenheit systemleistungsbedingter Zufriedenheitsunterschiede ergibt sich aus dem im Kontext der Benutzerleistung identifizierten systembedingten Anpassungseffekt der Relevanzwahrnehmung. Die Anwendung restriktiver bzw. weniger strenger Relevanzkriterien wirkt sich im Umkehrschluss auf die wahrgenommene Qualität des Suchsystems aus, die den objektiven Systemunterschied überlagert. Im Fall des besseren Systems führt dies zu einer Verminderung, im Fall des schlechteren Systems hingegen zu einer Zunahme der wahrgenommenen Systemqualität, wodurch der tatsächlich wahrgenommene Systemunterschied zwischen den beiden Untersuchungsgruppen geringer ausfällt. Im Ergebnis kann für viele der Zufriedenheitsindikatoren somit kein direkter Einfluss der Systemqualität auf die Benutzerzufriedenheit nachgewiesen werden. Zurückkommend auf die Eignung des EUCS-Instruments, lässt sich daraus ableiten, dass insbesondere der von den Benutzern wahrgenommene Sucherfolg im Kontext experimenteller Studien zum Informationssuchverhalten bei der Zufriedenheitsmessung stärker gewichtet werden sollte, da der im Rahmen der Interaktion mit dem System stattfindende Anpassungseffekt die Wahrnehmung der tatsächlichen Systemqualität verhindert. Der erwartungsbedingte Anpassungseffekt hingegen wirkt sich unabhängig vom objektiv vorhandenen Systemunterschied auf die wahrgenommene Systemqualität aus, was konsistent mit dem beschriebenen starken Einfluss der Erwartungshaltung auf das Zufriedenheitsurteil ist.

Zusammenfassend lässt sich festhalten, dass Benutzer zwar grundsätzlich in der Lage zu sein scheinen, Qualitätsunterschiede in Bezug auf die Systemleistung wahrzunehmen und diese anhand von Zufriedenheitsurteilen zu artikulieren. Im Zusammenspiel mit den beschriebenen Anpassungseffekten der Relevanzwahrnehmung überlagert die subjektiv wahrgenommene Systemleistung jedoch die objektiv vorhandenen Qualitätsunterschiede, wodurch sich nur für wenige Zufriedenheitsindikatoren eine signifikante Systemabhängigkeit einstellt.

8.2.2. Im Zuge der Untersuchungen gewonnene Einzelerkenntnisse zu FF2

Neben den über die Gesamtdaten gesicherten Ergebnissen der zweiten Forschungsfrage, liegen in dieser Arbeit auch im Zuge der Untersuchungen gewonnene Einzelerkenntnisse vor, die in den folgenden beiden Abschnitten zusammengefasst werden. Da es sich hierbei um weniger gesicherte Befunde handelt, muss hinsichtlich dieser Resultate jedoch darauf hingewiesen werden, dass weiterer Forschungsbedarf besteht, um die vermuteten theoretischen Zusammenhänge empirisch

weiter zu validieren.

8.2.2.1. Beeinflussung der Benutzerleistung durch die Systemqualität

Die verfügbaren Daten erlauben für den untersuchten Zusammenhang zwischen Systemqualität und Benutzerleistung weitere Schlussfolgerungen, die zu einem genaueren Verständnis der in den vorherigen Abschnitten beschriebenen Effekte beitragen. Im Wesentlichen ergeben sich drei Anhaltspunkte, die aufzeigen, in welcher Weise die beobachteten Anpassungseffekte der Relevanzbeurteilung im Suchkontext zustande kommen. Während der erste Anhaltspunkt in der im Rahmen des dritten Experiments zur Anwendung kommenden feineren Relevanzskala begründet liegt, ergibt sich der zweite Anhaltspunkt aus dem verwendeten stichprobendifferenzierenden Auswertungskonzept. Der dritte Anhaltspunkt schließlich bezieht die im dritten Experiment beobachteten Wechselwirkungen der präsentierten Systemqualität mit der über die Instruktion erhaltenen Erwartungsmanipulation in die Betrachtung mit ein.

Die im dritten Experiment zum Einsatz kommende 4-stufige Relevanzskala erlaubt zum einen einen differenzierteren Blick auf die Übertragbarkeit systemorientierter Evaluierungsergebnisse auf den Anwendungsfall. Zum anderen ermöglicht sie eine detailliertere Analyse der Anpassungseffekte der Relevanzwahrnehmung. Im vorliegenden Fall kann damit gezeigt werden, dass die Bewertungsanpassungen vorrangig im Bereich der mittleren Relevanzkategorien stattzufinden scheinen, also für solche Dokumente, deren Relevanz weniger eindeutig zu bestimmen ist. Diese Beobachtung trifft sowohl auf die Anwendung der restriktiveren Bewertungsstrategien im Fall der höheren Systemleistung als auch auf die Zugrundelegung der weniger restriktiven Maßstäbe im Fall der höheren Erwartungshaltung zu. Aus diesem Zusammenhang lässt sich schließen, dass Benutzer ihre Relevanzkriterien vor allem dann anzupassen scheinen, wenn die betrachteten Informationsobjekte in den Grenzbereich der beiden binären Relevanzkategorien fallen und somit schwieriger zu bewerten sind. Damit lässt sich die Frage nach dem Auslöser des Prozesses der Relevanzwahrnehmungsanpassung auf einen Zustand der Unsicherheit zurückführen, in dem Benutzer empfänglicher auf zusätzliche Hinweisreize (wie z.B. Systemgüte u. Erwartung) reagieren. Mit Bezug auf die von Scholer und Turpin (2008) und Scholer et al. (2008) betrachteten individuellen Relevanzschwellenwerte könnte dies darauf hindeuten, dass besagter Schwellenwert nicht nur eine individuelle Komponente aufweist, sondern zumindest hinsichtlich Systemleistung und Erwartungshaltung auch kontextabhängig ist. Des Weiteren wird an diesen Resultaten auch der Mehrwert feiner abgestufter Relevanzskalen im Kontext der IR-Evaluierung deutlich, auf den auch schon von Spink und Greisdorf (2001) und Sormunen (2002) hingewiesen wird. Bisher unbeantwortet bleibt jedoch die Frage, welche situationsbedingten Gründe Benutzer dazu veranlassen, im einen Fall die erhaltene Erwartungshaltung und im anderen Fall die präsentierte Systemleistung als Auslöser für die Bestimmung der jeweiligen Relevanzkriterien heranzuziehen.

Analog zu den nur vereinzelt auftretenden Effekten der Systemgüte im Fall der Benutzerzufriedenheit, kann für die Benutzerleistung im Rahmen der hier durchgeführten Experimente nur ein geringer Einfluss der Erwartungshaltung nachgewiesen werden. Während die Erwartungsmanipulation im ersten Experiment zu keinerlei signifikanten Änderungen führt, zeigt sich in Experiment 2 zumindest für eine Variable (Pre-Click-Precision (PCP)) bei der zweiten Aufgabe in der Tendenz ein positiver Zusammenhang zwischen der Erwartungshaltung der Benutzer und ihrer Suchleistung. Im dritten Experiment schließlich treten Haupteffekte der Erwartungshaltung

zum überwiegenden Teil nur in der bereinigten Stichprobe SP_B auf. Dies unterstreicht zusätzlich, dass derartige Effekte schwer messbar sind, weil sie sich nicht nur unmittelbar auf den Sucherfolg auswirken, sondern, wie im Stand der Forschung dargelegt (vgl. Kap. 2), mit anderen Faktoren wie bspw. Suchexpertise und Domänenwissen zusammenwirken können. Im Vergleich zwischen den beiden Teilstichproben entfällt mit 79% der Hauptanteil der in SP_B ausgeschlossenen Fälle auf Testpersonen, die unspezifische Suchbegriffe, häufig im Kontext der Wiki-Suchaufgabe, verwenden. Dies legt den Schluss nahe, dass die auffällige Abwesenheit erwartungsbedingter Anpassungseffekte in SP_A mit einem geringeren Vorwissen der Probanden verknüpft sein könnte. Ließe sich zunächst vielleicht vermuten, dass Nutzer sich gerade in Situationen, in welchen das eigene Vorwissen nicht ausreicht, auf ihre Erwartung verlassen, scheint es vielmehr so zu sein, dass eine erwartungsgesteuerte Adaption der Relevanzkriterien gerade ein gewisses Maß an Vorwissen bei den Testteilnehmern voraussetzt. Eine solche Interpretation steht im Einklang mit den in Abschnitt 2.1.1.3 berichteten Ergebnissen von Cox und Fisher (2004) und Vakkari und Hakala (2000), die zeigen, wie sich Erwartungen in Abhängigkeit von Suchexpertise und Domänenwissen verändern. In diesem Sinne scheint der systembedingte Anpassungseffekt hingegen weniger stark vom Vorwissen der Teilnehmer abhängig zu sein, da er stabil in SP_A und SP_B nachweisbar ist. Darüber hinaus transformiert sich für einige Benutzerleistungsmaße ein in der bereinigten Stichprobe SP_B signifikanter erwartungsbedingter Anpassungseffekt in der weniger bereinigten Stichprobe SP_A in ein systembedingtes Anpassungsverhalten. Bezüglich der im Rahmen der feineren Relevanzskala gestellten Frage nach den konkreten Beweggründen für das eine oder das andere Verhalten, könnte demnach eine mögliche Erklärung lauten, dass erwartungsbezogene Hinweisreize insbesondere in Situationen herangezogen werden, in welchen das eigene Vorwissen ausreicht, um die zu treffende Relevanzentscheidung beurteilen zu können. Demgegenüber scheint der Rückgriff auf systemleistungsbezogene Hinweisreize im Vergleich weniger von dem Vorwissen der Benutzer abzuhängen.

Die im Kontext von Experiment 3 nachgewiesenen Wechselwirkungen zwischen Systemqualität und Erwartungshaltung in Bezug auf die richtig bzw. falsch als eher irrelevant bewerteten Dokumente leisten einen weiteren Beitrag zu einem besseren Verständnis der beiden Anpassungseffekte. Es zeigt sich, dass die Anwendung einer systemleistungsbedingten strengeren bzw. weniger restriktiven Bewertungsstrategie am deutlichsten bei einer hohen Erwartungshaltung zu Tage tritt, während bei niedriger Erwartungshaltung nur ein geringer Unterschied zu beobachten ist. Dieser Zusammenhang lässt sich auch im zweiten Experiment für das Benutzerleistungsmaß $M03$ sowie teilweise auch für die dort durchgeführten Kovarianzanalysen bestätigen, wobei für ein Leistungsmaß jedoch auch eine Verstärkung des Anpassungseffekts bei niedriger Erwartungshaltung auftritt. Aus der Perspektive der erwartungsbezogenen Relevanzanpassung hingegen scheint sich der Effekt in Abhängigkeit der Systemleistung tendenziell umzukehren: Während bei niedriger Systemleistung eine hohe Erwartungshaltung zu einer weniger strengen Relevanzbewertung im Vergleich zur niedrigen Erwartung führt, ist im Fall des besseren Systems tendenziell der umgekehrte Effekt zu beobachten. Es zeigt sich somit, dass es potentiell hinsichtlich der beiden identifizierten Anpassungseffekte zu gegenseitigen Wechselwirkungen kommen kann, deren genaue gegenseitige Abhängigkeit jedoch in weiteren Studien geklärt werden muss. Die signifikante Interaktion zwischen Systemgüte und Erwartungshaltung, die sich im dritten Ex-

periment für die durchschnittliche Betrachtungszeit der als irrelevant bewerteten Dokumente (Z05) nachweisen lässt, vermittelt darüber hinaus einen weiteren Einblick in die weiter oben beschriebene Schwierigkeit, bei der Nutzung des besseren Systems Dokumente als irrelevant zu verwerfen: Es zeigt sich, dass hohe Erwartungen im Fall des besseren Systems zu längeren Betrachtungszeiten führen, während im Fall des schlechteren Systems kein Unterschied in Bezug auf die durchschnittlichen Lesezeiten der Probanden zu beobachten ist. Die Anwendung der strengeren Relevanzkriterien im Kontext der besseren Systemleistung und hoher Erwartungshaltung scheint somit mit einem erhöhten Aufwand durch die Probanden verbunden zu sein. Unter der Annahme, dass zusätzliche Hinweisreize insbesondere in schwierigen Entscheidungssituationen herangezogen werden, könnten die beobachteten Wechselwirkungen darauf hindeuten, dass im Zusammenhang mit einer hohen Erwartungshaltung dieser Hinweisreiz allein nicht ausreicht, um die bestehende Unsicherheit bezüglich der wahrgenommenen Relevanz zu überwinden, sodass in diesem Fall beide zur Verfügung stehenden Hinweisreize in die Relevanzbeurteilung einfließen.

In der Gesamtschau unterstreichen die in diesem Abschnitt zusammengefassten Befunde die funktionale Bedeutung von Benutzererwartungen im Suchprozess. Zwar scheint die Erwartungshaltung insgesamt gesehen nur einen mäßigen Einfluss auf die Benutzerleistung auszuüben, die betrachteten Wechselwirkungseffekte bieten jedoch die Grundlage für ein tiefergreifendes Verständnis der hinsichtlich der Relevanzbeurteilung identifizierten Anpassungseffekte, an welchen die Vorerwartung der Benutzer maßgeblich beteiligt zu sein scheint. Überdies legen die in diesem Abschnitt diskutierten Befunde einen engen Zusammenhang zwischen dem situativen Kontext und dem Auftreten der beschriebenen Anpassungsmechanismen nahe, demzufolge die beobachteten Effekte als Strategien zur Herstellung von Kohärenz verstanden werden können. Die Integration des Benutzers in den Evaluierungsprozess geht also in natürlicher Weise mit einer Erweiterung des Relevanzbegriffs einher, die über den Begriff der thematischen Relevanz hinaus zu der in der IIR-Theorie entwickelten Auffassung von einer situativen Relevanz führt (Saracevic, 1996; Borlund, 2003a).

8.2.2.2. Beeinflussung der Benutzerzufriedenheit durch die Systemqualität

Ein interessantes Einzelergebnis der Benutzerzufriedenheit betrifft die aktive Beteiligung des Benutzers am Suchprozess. Die in Abschnitt 8.2.1.2 dargestellten Ergebnisse deuten darauf hin, dass die subjektiv wahrgenommene Zufriedenheit nicht maßgeblich durch die objektiv messbare Systemqualität beeinflusst wird. Eine mögliche Erklärung dieses Verhaltens könnte darin liegen, dass Benutzer neben objektiven Gütekriterien auch ihren eigenen Beitrag am Suchergebnis als solchen erkennen und bei ihrer Zufriedenheitsbeurteilung berücksichtigen. Dieser Effekt könnte somit dazu beitragen, dass beide Suchmaschinen vergleichbare Zufriedenheitswerte erhalten. Um diese Vermutung zu prüfen, werden im Rahmen der Zufriedenheitsauswertung des dritten Experiments weitere differenzierende Analysen auf Grundlage leistungsbezogener Messindikatoren vorgenommen, die ausgewählte Aspekte des Aufwandes, der Effektivität sowie der Relevanzwahrnehmung der Testpersonen berücksichtigen. Die entsprechenden Kovarianzanalysen zeigen, wie vermutet, dass das Herauspartialisieren der eigenen Suchleistung den Einfluss der Systemqualität auf die Benutzerzufriedenheit stärker hervortreten lässt. Im Wesentlichen äußert sich dies in dem Hinzukommen vormals nicht signifikanter Systemeffekte oder dem Wegfall zuvor signifikanter Erwartungsabhängigkeiten. Obwohl dieses Verhalten vereinzelt auch für die

betrachteten Effektivitäts- und Aufwandsmaße zu beobachten ist, treten diese Effekte am deutlichsten im Kontext der wahrgenommenen Relevanz zu Tage und können wiederum im Sinne der in dieser Arbeit nachgewiesenen Anpassungsmechanismen ausgelegt werden: Hinsichtlich der Erwartungshaltung kann im Rahmen der Hauptanalyse gezeigt werden, dass eine höhere Erwartung zu einer weniger restriktiven Relevanzbewertung führt, welche das Zufriedenheitsurteil positiv beeinflusst (vgl. Abschn. 8.1.1). Insofern erscheint es nur folgerichtig, dass beim Herausrechnen dieses Einflusses aus der Benutzerzufriedenheit auch der dazugehörige Erwartungseffekt verschwindet. Auf ähnliche Weise lässt sich auch das Hinzukommen signifikanter Systemeffekte verstehen. Diesbezüglich ergibt die Hauptanalyse, dass eine höhere Systemleistung zu einer restriktiveren Relevanzbewertung führt, welche sich negativ auf das Zufriedenheitsurteil auswirkt (vgl. Abschn. 8.2.1.2). Im Ergebnis nehmen die Probanden beide Systeme als gleich gut wahr, was in dem geringen Einfluss der Systemgüte auf die Benutzerzufriedenheit zum Ausdruck kommt. Korrigiert man nun jedoch das Zufriedenheitsurteil gerade um diesen Anteil der wahrgenommenen Systemleistung, so tritt der tatsächliche Systemunterschied zu Tage und die Probanden zeigen sich mit dem besseren System zufriedener. Es zeigt sich also in beiden Fällen, dass die von den Testteilnehmern wahrgenommene und durch die jeweiligen Anpassungseffekte korrigierte Systemleistung einen unmittelbaren Einfluss auf das Zufriedenheitsurteil ausübt.

Die im Zuge der Bearbeitung der ersten Forschungsfrage entstandene Hypothese, die Wahrnehmung der eigenen Suchleistung könnte im Kontext der Informationssuche dazu beitragen, dass die Vorhersagen des C/D-Paradigmas nicht erfüllt werden, lässt sich somit nur zum Teil bestätigen (vgl. Abschn. 8.1.1). Zwar wird aus den Ergebnissen der Kovarianzanalysen deutlich, dass das systembedingte Anpassungsverhalten in Bezug auf die Relevanzwahrnehmung für den mangelnden Einfluss der Systemgüte auf die Benutzerzufriedenheit verantwortlich zu sein scheint, doch bleiben Detailfragen ungeklärt: So stellt sich bspw. die Frage, ob gerade das Herauspartialisieren der leistungsbezogenen Kontrollvariablen, das die Sichtbarkeit des Systemeffekts erhöht, nicht auch gleichzeitig den Einfluss der Erwartungshaltung so stark unterdrückt, dass das C/D-Paradigma nicht mehr sichtbar ist. Der Wegfall erwartungsbezogener Haupteffekte im Rahmen der Kovarianzanalyse spricht zunächst für diese These. Auf der anderen Seite treten in diesem Analyseschritt jedoch auch signifikante Wechselwirkungen auf, die weiterhin im Sinne eines erwartungsabhängigen Systemleistungsanpassungseffekts interpretiert werden können: Unterschiede in Bezug auf die Benutzerzufriedenheit sind nur bei Abwesenheit des systembedingten Anpassungseffekts, also im Fall der niedrigen Erwartungshaltung, zu beobachten.

Vor dem Hintergrund dieser signifikanten Wechselwirkungen kann darüber hinaus auch noch einmal auf grundlegender Ebene die Wirkung der Erwartungsmanipulation auf die Erwartungshaltung der Testteilnehmer diskutiert werden. Dabei fällt zunächst auf, dass das im Rahmen der Kovarianzanalyse erhaltene Interaktionsdiagramm (vgl. Abb. 8.1 (a)) keinem der in Abschnitt 4.2.1.2 diskutierten Szenarien entspricht. Insbesondere scheinen die Ergebnisse nicht auf eine Absenkung der prädiktiven und normativen Erwartungen im Fall der niedrigen Erwartungshaltung hinzudeuten (vgl. Abb. 4.3). Wird nämlich allein der in der Kundenzufriedenheitsforschung angenommene Soll-Ist-Vergleich zur Entstehung von Zufriedenheit zugrunde gelegt, deutet der beobachtete Zusammenhang zwischen Systemgüte, Erwartungshaltung und Benutzerzufriedenheit vielmehr auf eine Ausdehnung der Toleranzzone durch eine Verringerung der prädiktiven

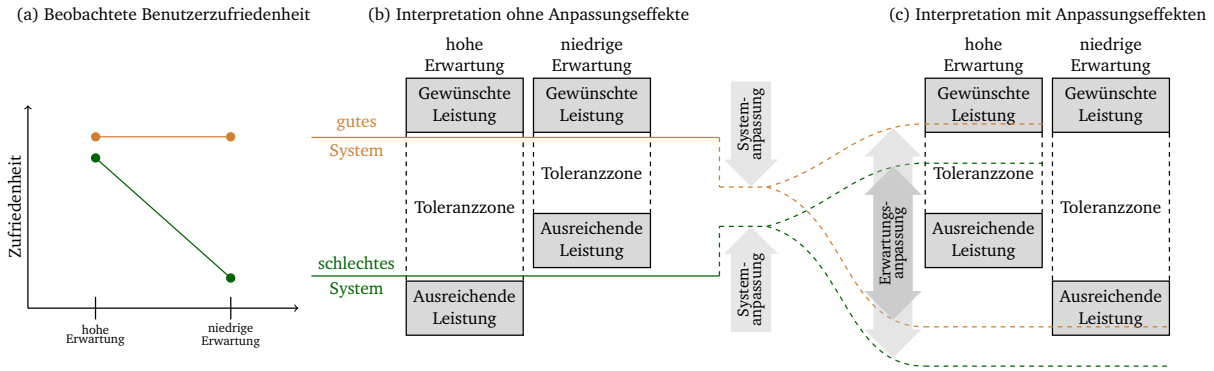


Abb. 8.1.: Wirkungsweise der Erwartungsmanipulation.

Erwartungen im Fall der hohen Erwartungshaltung hin (vgl. Abb. 8.1 (b)). Dies wäre jedoch der eigentlich intendierten Wirkung der Erwartungsmanipulation diametral entgegengesetzt. Allerdings lässt diese Argumentation die beschriebenen Anpassungseffekt der Relevanzwahrnehmung in Bezug auf Systemleistung und Erwartungshaltung außer Acht. Eine erweiterte Interpretation, die diese Effekt mit berücksichtigt und auf einer erweiterten Toleranzzone im Fall der niedrigen Erwartungshaltung beruht, ist hingegen in Abbildung 8.1 (c) dargestellt: Während der Systemanpassungseffekt den durch die Teilnehmer perzipierten Systemleistungsunterschied reduziert, führen die beiden Erwartungshaltungen zu einer erhöhten bzw. verringerten wahrgenommenen Systemgüte, die jedoch aufgrund der voneinander abweichenden Toleranzzonen zu den beobachteten unterschiedlichen Zufriedenheitsreaktionen bei hoher und niedriger Erwartungshaltung führen. In diesem Sinne kann die beobachtete Interaktion zwischen Systemgüte und Erwartungshaltung also als Wechselwirkung zwischen den Anpassungseffekten auf der einen und der durch die Erwartungsmanipulation geänderten Toleranzzonen auf der anderen Seite begriffen werden. Diese Sichtweise verdeutlicht noch einmal die Bedeutung der Wahrnehmung der eigenen Suchleistung für den Entstehungsprozess der Benutzerzufriedenheit im Kontext des IR.

Zurückkommend auf die Gültigkeit des C/D-Paradigmas für IR-Nutzerstudien, stellt sich außerdem die Frage, ob der gewählte relative Systemunterschied zwischen den beiden Suchsystemen stark genug ausfällt, um die Zufriedenheitsreaktion im Zusammenwirken mit der Erwartungshaltung und anderen Faktoren maßgeblich zu beeinflussen. Dieser Punkt berührt darüber hinaus auch die allgemeine Fragestellung, wie stark die Auswirkungen von Systemleistungsunterschieden im IR-Kontext maximal ausfallen können und ob die damit verbundenen Leistungsverbesserungen ausreichen, um das Auftreten diskonfirmatorischer Effekte zu rechtfertigen. Positiv ist jedoch hervorzuheben, dass die durch das C/D-Paradigma postulierte Annahme einer höheren Zufriedenheit als Ergebnis einer positiven Diskonfirmation in Bezug auf die Systemgüte durch die im Rahmen der Kovarianzanalysen auftretenden Interaktionseffekte bestätigt wird. Damit gewinnt die Untersuchung von Erwartungseinflüssen im Kontext experimenteller Studien zum Informationssuchverhalten eine neue Dimension: Erwartungen werden nicht mehr nur als Ausdruck von individuellen Bedürfnissen oder als Folge eigener Erfahrungen angesehen, sondern als Träger relevanter Information für die wechselseitige Interpretation suchbezogener Handlungen begriffen.

Zusammenfassend leisten die Ergebnisse der Kovarianzanalyse damit einen weiteren Beitrag zu einem besseren Verständnis benutzerzentrierter Evaluierungsprozesse. Es wird deutlich, dass die beobachteten Fakten und Zusammenhänge mitunter sehr komplex sein und erst durch ihre Betrachtung aus unterschiedlichen Perspektiven besser verstanden werden können. Bezogen auf die konkreten Ergebnisse dieser Arbeit kann darüber hinaus belegt werden, dass der Einfluss der Systemleistung auf die Benutzerzufriedenheit durch den systembedingten Anpassungseffekt der Relevanzwahrnehmung überlagert wird.

8.3. Forschungsfrage FF3: Dynamik der Suchergebniswahrnehmung

Nachdem sich die ersten beiden Forschungsfragen auf eine statische Analyse der Wahrnehmung und Bewertung von Suchergebnissen beschränken, beschäftigt sich die dritte Forschungsfrage schwerpunktmäßig mit der Veränderung der Wahrnehmung im Verlauf des interaktiven Suchprozesses. Aspekte, die in diesem Zusammenhang interessieren, betreffen z.B. die Stabilität der Wahrnehmung einer festen Systemgüte über mehrere Suchen und Aufgaben hinweg oder die Art und Weise, in welcher mit einer nicht zutreffenden Erwartungsmanipulation verfahren wird. Die wesentlichen Ergebnisse bezüglich der Dynamik von Benutzerleistung und Benutzerzufriedenheit werden im Folgenden dargestellt. Dabei ist zu beachten, dass diese Ergebnisse weniger aussagekräftig sind als bspw. die Ergebnisse zum Einfluss der Systemqualität, da die Dynamik des Suchprozesses ausschließlich bei der Auswertung des dritten Experiments mit in die Betrachtung einbezogen wird.

8.3.1. Beeinflussung der Benutzerleistung

Der folgende Abschnitt geht genauer auf die dynamische Entwicklung der Benutzerleistung im Rahmen des dritten Experiments ein. Damit wird insbesondere einer interaktiven Suchrealität Rechnung getragen, in der der Nutzer dynamisch und flexibel auf wechselnde situative Anforderungen reagiert. Diese Sichtweise steht auch im Einklang mit dem in Abschnitt 3.1.1 erläuterten Begriff der situativen Relevanz, wie er sich im Kontext des benutzerorientierten Evaluierungsansatz herausgebildet hat (Wilson, 1973; Schamber et al., 1990; Borlund u. Ingwersen, 1997). Darüber hinaus ermöglicht die dynamische Analyse ein besseres Verständnis einzelner im Zuge der statischen Untersuchung der ersten beiden Forschungsfragen gewonnenen Erkenntnisse. Eine natürliche Herangehensweise, um die dynamische Entwicklung der Benutzerleistung zu ergründen, besteht darin, zunächst mögliche leistungsbezogene Lern- oder Ermüdungseffekte zu identifizieren und anschließend ihre Auswirkungen zu bewerten. Neben allgemeinen Lerneffekten, die sich etwa in niedrigeren Verweildauern, geringeren Auffindezeiten für relevante Dokumente sowie einer höheren Übereinstimmung zwischen Benutzer- und Jurorenurteil äußern, kann so bspw. die schon bei der Analyse der Mittelwerte gemachte Beobachtung eines Zusammenhangs zwischen precisionorientierten Benutzerleistungsmaßen und der Systemleistung weiter präzisiert werden. Die Ergebnisse aller drei Untersuchungen zeigen diesbezüglich, dass eine bessere Systemleistung in den meisten Fällen mit einer höheren Benutzerleistung assoziiert ist. Zugleich stellt sich jedoch heraus, dass eine höhere Systemleistung teilweise mit einer restriktiveren Relevanzbewertung einhergeht und Nutzer des schlechteren Systems ihren Leistungsrückstand in diesen Fällen umkehren können. Die dynamische Analyse der Suchergeb-

niswahrnehmung erweitert diese Befunde nochmals, indem sie das Bewertungsverhalten nicht nur auf Ebene der Mittelwerte, sondern auch auf Ebene der einzelnen Suchaufgaben untersucht. So stellt sich für einzelne Precisionmaße, wie bspw. die Pre-Click-Precision (PCP), heraus, dass die beobachtete positive Korrelation zwischen Benutzerleistung und Systemqualität erst bei der letzten der drei Suchaufgaben wirksam wird: Die Nutzer des besseren Systems benötigen also eine gewisse Trainingsphase, bevor sie in der Lage sind, von der größeren Anzahl relevanter Dokumente in der Ergebnisliste zu profitieren und diese schon auf Ebene der Dokumentenselektion zu erkennen. Ein solches Verhalten kann als Hürde für den Wechsel zu einem alternativen Suchsystem angesehen werden, da ein Profitieren von einer möglicherweise höheren Suchleistung voraussetzt, dass der Nutzer die relevanten Dokumente auch erkennen kann. Da dies erst nach einer gewissen Einarbeitungszeit der Fall zu sein scheint, könnte dies den Nutzer veranlassen, wegen der zunächst ausbleibenden Verbesserung zu dem ursprünglichen Suchsystem zurückzukehren.

Darüber hinaus können auch für das systembedingte Anpassungsverhalten bei einzelnen Benutzerleistungsmaßen dynamische Effekte beobachtet werden. Dabei zeigt sich zum einen, dass, wie im statischen Fall, auch die dynamische Änderung des systembedingten Anpassungseffekts durch die Erwartungshaltung moderiert wird: Während im Fall der niedrigen Erwartungshaltung die Unterschiede hinsichtlich der Systemleistung im Wesentlichen zeitlich konstant sind, verstärken sich die Unterschiede bei hoher Erwartungshaltung im Zeitverlauf. Eine zweite Beobachtung betrifft Unterschiede in der zeitlichen Entwicklung des Anpassungseffekts innerhalb des Suchverlaufs der einzelnen Aufgaben, was auf eine noch vielfältigere dynamische Abhängigkeit hindeutet. Genauer scheint sich der systembedingte Anpassungseffekt in Bezug auf die Bewertung relevanter Dokumente bei der ersten Suche über die Zeit zu reduzieren bzw. sogar umzukehren, während für die letzte Suche und über den gesamten Suchverlauf hinweg, der Effekt stabil nachweisbar bleibt. Dies zeigt einmal mehr, dass die tatsächliche Dynamik der individuellen Relevanzbeurteilung noch weitaus komplexer ausfallen kann. Um hier jedoch genauere und verlässlichere Aussagen treffen zu können, bedarf es weiterer Untersuchungen.

Als Ergebnis lässt sich abschließend festhalten, dass es in benutzerorientierten Evaluierungskontexten bedeutsam ist, auch dynamische Faktoren in die Betrachtung einzubeziehen, da auf diese Weise u.U. weitere Anhaltspunkte für die analytische Beurteilung gewonnen werden können. Wie gezeigt werden kann, trifft dies insbesondere auch auf die Untersuchung von Erwartungseinflüssen zu. Zwar sollten die soeben beschriebenen Ergebnisse mit Vorsicht interpretiert werden, da die Voraussetzungen zur Durchführung einer klassischen Varianzanalyse in den meisten Fällen nicht erfüllt sind. Dennoch legt die Analyse nahe, dass auch die dynamische Änderung der durch die Systemqualität hervorgerufenen Anpassung der Relevanzwahrnehmung durch die Erwartungshaltung moderiert wird.

8.3.2. Beeinflussung der Benutzerzufriedenheit

Nachdem im vorangegangenen Abschnitt Effekte diskutiert werden, die auf eine dynamische Entwicklung der Benutzerleistung hindeuten, wird im Folgenden ein interessantes Einzelergebnis zur Rolle der Dynamik für die Evaluation der Benutzerzufriedenheit berichtet. Im Gegensatz zur Analyse der Benutzerleistung lassen sich im Fall der Benutzerzufriedenheit keine Interaktionen feststellen und auch der im Rahmen der statischen Analysen zumindest vereinzelt sichtbare Ein-

fluss der Systemgüte verschwindet im Kontext der dynamischen Zufriedenheitsanalyse zunächst vollständig. Um dennoch einen ersten Eindruck von der dynamischen Abhängigkeit der Nutzerzufriedenheit von Systemgüte und Erwartung zu erhalten, werden in diesem Fall auch Zufriedenheitsvariablen mit in die Betrachtung einbezogen, für die sich zumindest einmalig über die fünf Stichproben hinweg signifikante Wechselwirkungseffekte zeigen. Zwar führt die geringere Stabilität dieser Effekte wiederum zu einer eingeschränkten Generalisierbarkeit, jedoch lassen sich auf diese Weise erste Trends erkennen. Konkret gibt die Analyse der Zufriedenheitsdynamik in diesem Fall Anlass zu der Vermutung, dass der Erwartungseinfluss – entgegen der vierten Forschungshypothese des dritten Experiments – nicht abklingt, sondern sich eher verstärkt. Gleichzeitig scheint der Einfluss der Systemqualität tendenziell nur im Falle einer niedrigen Erwartungshaltung sichtbar zu werden, also genau dann, wenn nach Abschnitt 8.2.2.1 keine Anpassung der Relevanzkriterien stattfindet. In diesem Fall können die Probanden den Systemunterschied also wahrnehmen und eine anfänglich negative Erwartungshaltung scheint in Kombination mit dem schlechteren System über die Zeit zu einem immer negativeren Zufriedenheitsurteil zu führen, wohingegen die Zufriedenheit mit dem besseren System im Wesentlichen zeitlich konstant bleibt. Kommt es hingegen zu einem systembedingten Anpassungseffekt (bei hoher Erwartungshaltung) können die Probanden den Systemunterschied nicht wahrnehmen und in Folge dessen findet auch keine Zunahme bzw. Abnahme der Zufriedenheit in Abhängigkeit von der Systemgüte statt.

Zusammenfassend lässt sich damit auch für die Benutzerzufriedenheit feststellen, dass dynamische Faktoren die Suchergebniswahrnehmung beeinflussen können. So finden sich Hinweise, dass der Erwartungseinfluss nicht, wie ursprünglich vermutet, mit der Zeit schwindet, sondern sich unter gewissen Umständen sogar intensiviert. Allerdings gilt dieses Ergebnis nur für einen Teil der untersuchten Stichprobe und sollte deshalb zunächst durch weitere Studien bestätigt werden.

8.4. Fazit

Wie in den vorangegangenen Abschnitten aus einer vergleichenden Perspektive dargestellt, erlauben die im Rahmen dieser Arbeit durchgeführten Nutzerstudien Aussagen im Hinblick auf alle drei untersuchten Forschungsfragen. Dabei zeigt sich zum einen, dass die Vorhersagen des C/D-Paradigmas im Kontext des IR nicht vollumfänglich sichtbar werden (FF1), während sich andererseits jedoch Hinweise sowohl für einen Einfluss von Systemleistung und Erwartungshaltung auf das Nutzerverhalten (FF2) wie auch auf dynamische Komponenten der System-Benutzer-Interaktion (FF3) ergeben. Genauer können in der Gesamtschau drei zentrale Ergebnisse festgehalten werden.

Zunächst kann in Übereinstimmung mit früheren Untersuchungen bestätigt werden, dass Suchmaschinennutzer in der Lage sind, systembedingte Qualitätsunterschiede in Bezug auf die Vollständigkeit der Recherche (Recall) auszugleichen. Es wird deutlich, dass dieser recallbezogene Kompensationseffekt äußerst stabil ist und Nutzer somit unabhängig von der Systemqualität im Stande sind, ihr Informationsbedürfnis zu befriedigen. Gleichzeitig kann dieser Effekt als erster Hinweis auf ein situativ angepasstes Nutzerverhalten im Kontext des Suchprozesses gedeutet werden.

Unterschiede in der Benutzerleistung hingegen lassen sich vor allem hinsichtlich der Genauig-

keit der Suchergebnisse (Precision) beobachten. Hier kann eine Anpassung der Relevanzkriterien, die bei steigender Systemleistung zu strengeren Bewertungsmaßstäben führt, stabil über alle drei Nutzerstudien hinweg, nachgewiesen werden. Hervorzuheben sind in diesem Zusammenhang die im Rahmen dieser Arbeit eingeführten Imprecisionmaße, die sich im Fall der letzten beiden Experimente als besonders geeignet herausstellen, diese Unterschiede in der Relevanzwahrnehmung zu detektieren. Diese neue Klasse von Leistungsindikatoren stellt somit eine interessante Erweiterung des Kanons der standardmäßig in IIR-Studien verwendeten Benutzerleistungsmaße dar.

Darüber hinaus zeigen die Ergebnisse dieser Arbeit, wie eine kontrollierte Manipulation der Erwartungshaltung im Rahmen von IIR-Experimenten praktisch umgesetzt werden kann und zu messbaren Veränderungen des Nutzerverhaltens führt. Im Vergleich zeigt sich, dass die Systemgüte einen stärkeren Beitrag zur Erklärung der Benutzerleistung leistet, während Erwartungen einen größeren Einfluss auf die Benutzerzufriedenheit auszuüben scheinen. Allerdings kann darüber hinaus auch für die Erwartungshaltung ein signifikanter Einfluss auf die Relevanzwahrnehmung nachgewiesen werden, bei dem eine steigende Erwartung mit weniger restriktiven Relevanzkriterien einhergeht. Systemqualität und Erwartungshaltung stellen also tatsächlich einander ergänzende Faktoren dar, die bei der Untersuchung der Suchergebniswahrnehmung gleichermaßen berücksichtigt werden sollten.

Diese Hauptergebnisse werden durch weitere Einzelergebnisse der drei Nutzerstudien präzisiert, die sich jeweils auf spezielle Aspekte des gewählten Untersuchungsdesigns beziehen. So bietet bspw. die im dritten Experiment eingeführte 4-stufige Relevanzskala die Möglichkeit, die beobachteten Anpassungseffekte hinsichtlich der Relevanzbewertung von Dokumenten im Grenzbereich zwischen relevant und irrelevant zu verorten. In ähnlicher Weise wie bei der Berücksichtigung von Störfaktoren durch Kovarianzanalysen ergeben sich für 4-stufige Relevanzmaße darüber hinaus auch Wechselwirkungen zwischen Systemgüte und Erwartungshaltung, die weiteren Aufschluss über die Anpassung der Relevanzwahrnehmung liefern. Im Ergebnis ist der systembedingte Anpassungseffekt im Wesentlichen nur im Zusammenspiel mit der hohen Erwartungshaltung zu beobachten. Des Weiteren bietet der systembedingte Anpassungseffekt, indem er den wahrgenommenen Systemleistungsunterschied reduziert, einen Erklärungsansatz für die schwache Abhängigkeit der Benutzerzufriedenheit von der Systemgüte. Diese Interpretation wird durch die Tatsache gestützt, dass das Herauspartialisieren des Anpassungseffekts aus den Zufriedenheitsurteilen im Rahmen der Kovarianzanalyse die zu vermutende positive Korrelation zwischen Systemgüte und Benutzerzufriedenheit sichtbar macht. Diese Hypothese ist auch im Hinblick auf die beobachteten Wechselwirkungen zwischen Systemgüte und Erwartungshaltung im Kontext von Benutzerleistung und Benutzerzufriedenheit konsistent.

Weiterhin gibt die Berücksichtigung der zeitlichen Veränderung des Nutzerverhaltens erste Hinweise auf dynamische Abhängigkeiten von Systemgüten- und Erwartungseinfluss. Neben der Identifikation von Lern- und Ermüdungseffekten sind hier insbesondere die Zunahme von Erwartungs- bzw. Systemeffekten in Bezug auf die Anpassungseffekte zu nennen, sowie erste Hinweise auf dynamische Verhaltensänderungen innerhalb einzelner Suchsessions, die sich im Vergleich zwischen der ersten und letzten durchgeführten Suche manifestieren.

Im Anschluss an die diesem Kapitel zugrunde liegende vergleichende Perspektive auf die Ergeb-

nisse der durchgeführten Nutzerstudien erfolgt im nächsten Kapitel eine Einordnung der Befunde im Hinblick auf die ihnen zugrunde liegenden Gütekriterien, praktische Handlungsoptionen für die Weiterentwicklung von Suchsystemen sowie weiteren Forschungsbedarf.

9. Diskussion und Ausblick

Den Ausgangspunkt dieser Arbeit bildet die Annahme, dass Prognosen über den zu erwartenden Sucherfolg regulärer Bestandteil einer jeden Suchsituation sind. In diesem Sinne stellen Erwartungen einen Bezugsrahmen bereit, um die Konsequenzen für zukünftiges Verhalten besser einschätzen zu können und Unsicherheiten (z.B. bei der Bewertung von Informationsobjekten) zu reduzieren, indem auf Vorwissen zurückgegriffen wird. Während bisherige Forschungsarbeiten oftmals entweder die Übertragbarkeit systemorientierter Retrievalergebnisse auf den Anwendungsfall in den Blick nehmen (Hersh et al., 2000; Turpin u. Hersh, 2001; Allan et al., 2005; Turpin u. Scholer, 2006) oder den Zusammenhang zwischen Sucherfolg und Nutzungszufriedenheit auf Ebene des für die Befriedigung eines Informationsbedürfnisses zu erbringenden Aufwands analysieren (Al-Maskari u. Sanderson, 2010; Kelly et al., 2007; Xu u. Mease, 2009; Kiseleva et al., 2016; Luo et al., 2017), ist das primäre Ziel dieser Arbeit eine Kombination dieser beiden Perspektiven des Sucherfolgs, die darüber hinaus neben der individuellen Relevanzwahrnehmung einzelner Informationsobjekte den potentiellen Einfluss von Erwartungen auf die Suchzufriedenheit in den Fokus rückt. Davon ausgehend, dass beide Ansätze einander ergänzende Herangehensweisen darstellen, werden Benutzerleistung und Benutzerzufriedenheit innerhalb eines gemeinsamen experimentellen Forschungsdesigns untersucht. Neben der Frage, inwiefern systemleistungsbezogene Unterschiede die Wahrnehmungs- und Handlungsweisen von Suchmaschinennutzern beeinflussen, soll dabei insbesondere auch der Frage nachgegangen werden, ob das aus der Kundenzufriedenheitsforschung stammende C/D-Paradigma zur Entstehung von Zufriedenheit auch im Kontext des IR seine Gültigkeit behält. Zur Beantwortung dieser Forschungsfragen wird ein Untersuchungsdesign entworfen, mit dessen Hilfe das unter natürlichen Bedingungen nur schwer zugängliche Bewertungsverhalten von Informationssuchenden unter kontrollierten Bedingungen analysiert werden kann. Die Ergebnisse der auf diesem Untersuchungsdesign basierenden Experimente werden in Kapitel 8 ausführlich diskutiert. Das folgende Kapitel widmet sich der Interpretation und Bewertung der erhaltenen Befunde. Der erste Abschnitt betrachtet zunächst Stärken und Einschränkungen dieser Arbeit. Im zweiten Abschnitt werden Überlegungen und Möglichkeiten zur weiterführenden Untersuchung der hier behandelten und im Verlauf des Forschungsprojekts aufgetauchten Fragestellungen sowie zur praktischen Nutzung der erhaltenen Befunde skizziert.

9.1. Stärken und Einschränkungen dieser Arbeit

Zur Einschätzung der Stärken und Schwächen werden im Folgenden unterschiedliche Ebenen des forschungsmethodischen Vorgehens in den Blick genommen: der theoretische und empirische Beitrag dieser Arbeit, die Wahl und Operationalisierungen der erhobenen Konstrukte, die Gestaltung des zugrundeliegenden Forschungsprozesses sowie die Zuverlässigkeit der erhaltenen

Ergebnisse.

9.1.1. Zum theoretischen und empirischen Beitrag dieser Arbeit

Im Rahmen der vorliegenden Arbeit werden Systemqualität und Erwartung erstmals explizit gemeinsam als Determinanten eines erfolgreichen Suchprozesses untersucht. Die Ergebnisse bestätigen, dass sowohl Systemgüte als auch Erwartungshaltung eine bedeutende Rolle im Prozess der Informationssuche einnehmen. Insbesondere die systematische Untersuchung des komplexen Zusammenspiels zwischen Systemleistung und Erwartungshaltung stellt einen zentralen Beitrag der vorliegenden Arbeit dar. Durch die gemeinsame Untersuchung dieser beiden Variablen wird deutlich, dass Systemgüte und Benutzererwartung für sich genommen bereits bedeutsame Determinanten eines erfolgreichen Suchprozesses bilden, dass aber auch ein Zusammenwirken besteht. Es können zwei die Relevanzbeurteilung der Testteilnehmer betreffende Anpassungseffekte identifiziert werden, deren Auftreten Auswirkungen auf beide betrachteten Zielfaktoren hat, nämlich sowohl auf die wahrgenommene Effektivität aus Sicht der Probanden, also auch auf den objektiv bestimmbaren Benutzererfolg: Während die restriktivere Bewertungsstrategie durch Benutzer des besseren Systems im Fall des systemleistungsbedingten Anpassungsverhaltens dazu beitragen kann, dass Nutzer des schlechteren Systems ihren Leistungsrückstand umkehren können und die Effektivität des Testsystems auf einem vergleichbaren Niveau wahrnehmen wie Nutzer des besseren Systems, hat das positivere Bewertungsverhalten durch Probanden mit hohen Erwartungen im Fall des erwartungsbedingten Anpassungseffekts zur Folge, dass sich die erhaltene Erwartungsmanipulation im Zufriedenheitsurteil der Probanden widerspiegelt. Besondere Aufmerksamkeit verdient in diesem Zusammenhang auch die Untersuchung weiterer Kontextfaktoren wie z.B. der Detaillierungsgrad der Relevanzbeurteilung, das Vorwissen der Testteilnehmer oder die im Kontext des dritten Experiments nachgewiesenen Wechselwirkungen. Die zusätzliche Analyse dieser Faktoren trägt im Rahmen dieser Arbeit zu einem genaueren Verständnis der zuvor beschriebenen Anpassungseffekte bei. Es kann gezeigt werden, dass die besagten Bewertungsanpassungen vorrangig im Bereich der mittleren Relevanzkategorien stattzufinden scheinen, also im Hinblick auf Dokumente, deren Relevanz weniger eindeutig zu bestimmen ist. Auslöser des Prozesses der Relevanzwahrnehmungsanpassung scheint also ein Zustand der Unsicherheit zu sein, in dem Benutzer empfänglicher auf zusätzliche Hinweisreize reagieren. In Bezug auf den systemleistungsbedingten Anpassungseffekt scheint sich außerdem eine an die Voreinstellung gebundene Abhängigkeit abzuzeichnen, wonach eine Anpassung der Relevanzkriterien insbesondere im Zusammenhang mit der positiven Erwartungsmanipulation vorgenommen zu werden scheint. Der Einfluss der Erwartungshaltung auf die Benutzerleistung hingegen scheint in stärkerem Maße von einem ausreichenden Vorwissen der Benutzer abhängig zu sein. Bezüglich der Frage nach den konkreten Beweggründen für das eine oder das andere Verhalten, könnte demnach eine mögliche Erklärung lauten, dass erwartungsbezogene Hinweisreize insbesondere in Situationen herangezogen werden, in welchen das eigene Vorwissen ausreicht, um die zu treffende Relevanzentscheidung beurteilen zu können. Demgegenüber scheint der Rückgriff auf systemleistungsbezogene Hinweisreize im Vergleich weniger von dem Vorwissen als von der Erwartungshaltung der Benutzer abzuhängen.

Darüber hinaus wird mit dem C/D-Paradigma in dieser Arbeit der Versuch unternommen, ein theoretisches Modell aus der Kundenzufriedenheitsforschung auf den Kontext der Infor-

mationssuche zu übertragen, das die Zufriedenheitsreaktion als Ergebnis eines individuellen Soll-Ist-Vergleichs begreift. Dieses Modell bildet den theoretischen Rahmen für die durchgeführten Benutzerstudien. Das darauf aufbauend entwickelte Untersuchungsdesign stellt den empirischen Kern dieser Arbeit dar. Entgegen der üblichen Vorgehensweise bei experimentellen Studien zum Informationssuchverhalten, wird in den hier durchgeführten Untersuchungen ein Between-Subjects-Design zugrunde gelegt. Neben einem geringeren zeitlichen Aufwand für die Testpersonen, besteht ein wesentlicher Vorteil dieses Forschungsdesigns in der Verminderung möglicher Ausstrahlungseffekte, die durch die Konfrontation mit unterschiedlichen Variationen der unabhängigen Variablen entstehen können. Dabei zeigt sich, dass die Entstehung von Benutzerzufriedenheit zwar auch im IR-Kontext als Soll-Ist-Vergleich zwischen perzipierter Systemleistung und vorliegender Erwartungshaltung verstanden werden kann, die wahrgenommene Systemgüte selbst jedoch auch von der Nutzererwartung beeinflusst wird und somit nicht allein von der Qualität des Suchsystems abhängt. Dieser doppelte Einfluss der Erwartungshaltung führt im Ergebnis dazu, dass die Vorhersagen des C/D-Paradigmas nicht vollständig reproduziert, sondern durch den Einfluss der nachgewiesenen Anpassungseffekte modifiziert werden.

In Bezug auf das methodische Vorgehen ist das im Rahmen der letzten beiden Experimente gewählte stichprobendifferenzierende Auswertungskonzept hervorzuheben, welches erlaubt Aussagen auf einer allgemeineren Ebene zu treffen, da kritische Fälle nicht von vornherein von der Auswertung ausgeschlossen werden müssen. Im Sinne eines Top-down Vorgehens werden die Daten zunächst auf Ebene des weniger streng kontrollierten Gesamtdatensatz SP_A analysiert und erst im Anschluss überprüft, ob der Ausschluss kritischer Fälle (SP_B) das Untersuchungsergebnis signifikant verändert. Durch dieses parallele Vorgehen kann bspw. das im vorigen Abschnitt diskutierte Ergebnis einer aus dem Vorwissen resultierenden Reaktion der Benutzer auf unterschiedliche Hinweisreize begründet werden. Da das gewählte Vorgehen zur Stichprobenprüfung als konservativ, d.h. auf der sicheren Seite liegend, zu bezeichnen ist, sind im Fall des vollständig bereinigten Datensatzes SP_B allerdings nicht für alle Variablen ausreichend Fälle vorhanden, um eine statistische Auswertung zuzulassen. Dies führt dazu, dass bei der Auswertung zwischen mehr und weniger gesicherten Befunden unterschieden werden muss. Um noch validere Aussagen hinsichtlich der Belastbarkeit der mittels SP_A gewonnenen Ergebnisse treffen zu können, wäre eine zusätzliche Testung weiterer Probanden notwendig gewesen.

Ein weiterer wissenschaftlicher Beitrag dieser Arbeit liegt im Zugewinn des allgemeinen Verständnisses über die Dynamik der Suchergebniswahrnehmung in Abhängigkeit unterschiedlicher Systemqualitäten und Erwartungshaltungen. Zwar ist zu beachten, dass die im Zuge der dynamischen Analyse der Suchergebniswahrnehmung des dritten Experiments beobachteten Interaktionseffekte aufgrund der Verletzung der statistischen Voraussetzungen eher als Trend denn als gesicherte, generalisierbare Befunde anzusehen sind, dennoch bestätigen die Ergebnisse erneut, dass sowohl Systemgüte als auch Erwartungshaltung bedeutende Faktoren im dynamischen Prozess der Informationssuche darstellen. Im Kontext der Benutzerleistung ergeben sich weitere Anhaltspunkte, die die Annahme einer moderierenden Rolle der Erwartungshaltung in Bezug auf das systemleistungsbezogene Anpassungsverhalten der Relevanzbewertung bestärken und insbesondere die hohe Erwartungshaltung mit diesem Verhalten verknüpfen. Darüber hinaus lassen sich auch hinsichtlich der Benutzerzufriedenheit dynamische Entwicklungen beobachten,

die mit der Erwartungshaltung der Benutzer variieren und der eingangs geäußerten Vermutung entgegen stehen, wonach der Einfluss von Erwartungen mit der Zeit abnehmen sollte. Um diese Aspekte der Dynamik der Suchergebniswahrnehmung abschließend zu klären, sind jedoch weitere Studien notwendig.

9.1.2. Zur Wahl und Operationalisierung der untersuchten Variablen

Zur Modellierung der System-Benutzer-Interaktion werden in dieser Arbeit kognitive, motivationale, demographische, individuelle und situative Einflussfaktoren herangezogen. Da das gewählte Untersuchungsdesign nicht alle möglichen Einflussfaktoren auf den Sucherfolg beinhalten kann, werden Einflussfaktoren ausgewählt, von denen auf Basis des Stands der Forschung davon ausgegangen werden kann, dass sie einen entscheidenden Beitrag zum Nutzerverhalten leisten. Um das dieser Arbeit als theoretischer Rahmen zugrunde gelegte C/D-Paradigma untersuchen zu können, werden Systemqualität und Erwartungshaltung als unabhängige Variablen manipuliert. Weitere Einflussfaktoren wie z.B. Alter, Geschlecht, Motivation und Vorwissen gehen hingegen als Kovariaten in die Analysen ein. Auf Seite der untersuchten Zielfaktoren wird auf eine Kombination aus Verhaltens- bzw. Leistungsmaßen zur objektiven Bestimmung der individuellen Suchleistung und subjektiven Zufriedenheitsindikatoren zur Ermittlung des wahrgenommenen Sucherfolgs zurückgegriffen. Dieser mehrdimensionale Ansatz bietet folgende Vorteile: Er erfasst unterschiedliche Aspekte des Sucherfolgs und macht diese durch die Erfassung von Leistungs- und Zufriedenheitswerten miteinander vergleichbar. Es finden somit sowohl objektiv messbare als auch durch die Teilnehmer individuell wahrgenommene Kriterien Berücksichtigung, wobei gerade die individuelle Komponente in letzter Konsequenz über die Qualität eines IR-Systems entscheidet. Darüber hinaus wird durch die Kombination unabhängig voneinander gemessener Benutzerindikatoren die Gefahr verringert, mögliche Messfehler nicht zu erkennen. Der wichtigste Vorteil liegt jedoch darin, dass durch die gemeinsame Untersuchung von Systemleistung und Erwartungshaltung – ganz im Sinne des C/D-Paradigmas – auch das Zusammenwirken dieser beiden unabhängigen Variablen berücksichtigt werden kann.

9.1.2.1. Operationalisierung der unabhängigen Variablen der System-Benutzer-Interaktion

Die verfügbare Systemleistung und die erhaltene Erwartungsmanipulation stellen die unabhängigen Variablen des vorliegenden Forschungsdesigns dar. Bei der Auswahl der jeweiligen Operationalisierungsstrategien werden folgende Leitlinien und Erfordernisse berücksichtigt: Beide Variablen sollten in ihrer Variation der alltäglichen Erfahrung möglichst nahe kommen und gleichzeitig plausibel, bedeutungsvoll und überzeugend sein. Darüber hinaus muss in beiden Fällen die Manipulationsstärke groß genug ausfallen, um entsprechende Effekte hinsichtlich der abhängigen Variablen sichtbar machen zu können. Gleichzeitig sollten die variierten Unterschiede jedoch nicht zu stark ausfallen, um sicherzustellen, dass die beobachteten Effekte nicht nur auf die untersuchten extremen Variablenausprägungen begrenzt sind. Sowohl das Ziel, einen ausreichenden Systemunterschied zwischen gutem und schlechtem System zu realisieren, als auch das Ziel, eine angemessene Erwartungsmanipulation zu finden, werden erreicht. Für einen relativen Systemunterschied von $0{,}35\%$ bezüglich der AvP kann über alle drei Experimente hinweg ein mehrheitlich positiver Effekt auf die Benutzerleistung nachgewiesen werden. Dabei ist anzumerken, dass die verwendeten Rankinglisten bezüglich anderer Systemleistungsmaße, wie

bspw. BPref, eine gute Trennung der beiden Systemqualitätsstufen aufweisen. Gleichzeitig zeigt sich für recallorientierte Benutzerleistungsmaße ein kompensatorisches Verhalten, das jedoch im Einklang mit anderen Studien steht (Al-Maskari et al., 2008b; Turpin u. Hersh, 2001; Allan et al., 2005). Während der objektiv messbare Sucherfolg stark durch die Systemleistung beeinflusst wird, ergibt sich für den Zusammenhang zwischen wahrgenommenem Sucherfolg und Systemleistung nur eine schwache Abhängigkeit, was jedoch auf eine Überlagerung des Systemeinflusses durch den systembedingten Anpassungseffekt der Relevanzwahrnehmung zurückgeführt werden kann (vgl. Abschn. 8.2.2.2).

Im Fall der Erwartungsmanipulation erweist sich die innere Vorstellung eines Systemvergleichs als besonders geeignet, um den Testpersonen die Einnahme der betreffenden Erwartungshaltung zu erleichtern. Als gelungen zeigt sich diese Manipulationsmethode sowohl in Bezug auf die Benutzerleistung, bei der sie zu dem erwartungsbedingten Anpassungseffekt der Relevanzwahrnehmung führt, als auch hinsichtlich der Benutzerzufriedenheit, die im Zusammenhang mit einer positiven Voreinstellung höher ausfällt. Insgesamt kann somit im Rahmen dieser Arbeit demonstriert werden, dass eine kontrollierte Manipulation der Erwartungshaltung im Kontext von IR-Experimenten erfolgreich durchführbar ist.

9.1.2.2. Wahl der abhängigen Variablen der System-Benutzer-Interaktion

Bei der Auswahl der für die abhängigen Variablen relevanten Indikatoren wird – soweit möglich – ebenfalls auf etablierte und getestete Messinstrumente zurückgegriffen. Im Fall der Benutzerleistung liegen hierbei eine Vielzahl von Effektivitäts-, Effizienz- und Aufwandsmaßen aus verschiedensten Studien im Bereich experimenteller Benutzerforschung vor (vgl. Abschn. 4.2.2.2). Zusätzlich eröffnet die Einführung einer feiner abgestuften Relevanzskala im Zuge des dritten Experiments die Möglichkeit, detailliertere Analysen vorzunehmen, mit deren Hilfe auch Aussagen über die Richtung der in Bezug auf die Relevanzwahrnehmung identifizierten Anpassungseffekte möglich sind. Insgesamt werden damit im zweiten Experiment 86 und im dritten Experiment 212 unterschiedliche Benutzerleistungsmaße berücksichtigt. Durch die Einbeziehung einer derart hohen Anzahl unterschiedlicher Leistungsmaße kann eine breite Abdeckung des Konstruktinhalts gewährleistet werden. Die Verwendung der üblichen Standardmaße sichert dabei die Validität der Daten und stellt gleichzeitig die Vergleichbarkeit der Ergebnisse mit anderen Untersuchungen sicher. Eine möglichst umfassende Abdeckung des Konstruktinhalts zu gewährleisten, scheint auch vor dem Hintergrund der verschiedenen, teils widersprüchlichen Forschungsergebnisse zur Rolle der Systemleistung im Benutzerkontext sinnvoll, um die erhaltenen Ergebnisse angemessen interpretieren zu können (vgl. Abschn. 3.2.1). In diesem Zusammenhang ist auch der in den Experimenten 2 und 3 aufgezeigte Mehrwert der im Rahmen dieser Arbeit eingeführten Imprecisionmaße zu erwähnen. Diese Leistungsindikatoren, die anstelle der Übereinstimmungs- die Widerspruchstendenz zwischen Juroren und Testteilnehmern erfassen, tragen entscheidend zu einem besseren Verständnis der in dieser Arbeit identifizierten Anpassungsreaktionen bei.

Für die Erfassung der Benutzerzufriedenheit bietet sich das aus der Informationssystemforschung stammende EUCS-Instrument zur Nachnutzung an. Dabei sollte ein geeignetes Fragebogeninstrument im Wesentlichen folgende Anforderungen erfüllen: Es sollte methodisch abgesichert sein, relevante Dimensionen der Suchergebniszufriedenheit erfassen und hinsichtlich der beinhalteten Skalen ein gute Differenzierungsfähigkeit aufweisen. Schnell zeigt sich, dass

das EUCS-Instrument diese Anforderungen nur bedingt erfüllt, da wichtige Aspekte der Informationssuche zunächst unberücksichtigt bleiben. Eine wesentliche Ergänzung des Instruments stellt daher insbesondere die im Rahmen dieser Arbeit vorgenommene Erweiterung des Fragenkatalogs um die subjektive Bewertung des persönlichen Sucherfolgs dar. Insgesamt kann dieser methodische Teilaspekt der vorliegenden Arbeit als erster Schritt hin zu einem evaluierten Fragebogeninstrument zur Erfassung der Nutzerzufriedenheit im Kontext von IR-Studien angesehen werden.

Abschließend bleibt noch einmal festzuhalten, dass gerade die Kombination objektiver und subjektiver Benutzermaße zu einem tieferen Verständnis des beobachteten Nutzerverhaltens beitragen kann. Exemplarisch zeigt sich dies für die Nutzerzufriedenheit, bei der die Abwesenheit eines Einflusses der Systemgüte durch die Verringerung des wahrgenommenen Systemunterschieds aufgrund einer systembedingten Anpassung der Relevanzwahrnehmung erklärt werden kann. Die Tatsache, dass das Herauspartialisieren leistungsbezogener Messindikatoren den Einfluss der Systemqualität auf die Benutzerzufriedenheit stärker hervortreten lässt, unterstreicht den methodischen Mehrwert eines solchen Vorgehens zusätzlich.

9.1.3. Zur Gestaltung des Forschungsprozesses und des Untersuchungsdesigns

Der im Rahmen dieser Arbeit verfolgte Forschungsprozess beruht wesentlich auf der Kombination der drei experimentellen Nutzerstudien. Dabei erlaubt das iterative Vorgehen bei der Gestaltung des Untersuchungsdesigns einerseits erkannte Defizite auszugleichen und andererseits eine Kreuzvalidierung der erhaltenen Ergebnisse. Studie 1 basiert auf einem 2×2-Design, welches sich primär durch seine Einfachheit auszeichnet. Ausgehend von den Vorhersagen des C/D-Paradigmas wird dabei von drei Grundannahmen ausgegangen: 1. Probanden mit unrealistisch niedrigen Erwartungen erleben im Verlauf der Suche eine positive Diskonfirmation und sind infolgedessen mit dem Testsystem zufrieden. 2. Probanden mit realistischerweise niedrigen oder hohen Erwartungen fühlen sich während der Suche in ihren Erwartungen bestätigt und sind deshalb ebenfalls mit dem Testsystem zufrieden. 3. Probanden mit unrealistisch hohen Erwartungen hingegen erleben im Laufe der Suche eine negative Diskonfirmation und sind infolgedessen mit dem Testsystem unzufrieden. Dies führt dazu, dass im ersten Experiment nur vier Untersuchungsgruppen benötigt werden, was nicht nur die statistische Auswertung vereinfacht. Da darüber hinaus lediglich die über die drei Suchaufgaben hinweg gemittelten Benutzerleistungsmaße bzw. die am Anschluss an die gesamte Suchsession erhobenen Zufriedenheitsurteile in die Analyse eingehen, reduziert sich der Auswertungsaufwand zusätzlich. Das gewählte Untersuchungsdesign erfüllt damit zunächst alle notwendigen Voraussetzungen, um die Übertragbarkeit des C/D-Paradigmas auf den Kontext der Informationssuche untersuchen zu können. Kritisch anzumerken ist hingegen die nicht erfolgreich zu verlaufen erscheinende Erwartungsmanipulation. Aussagen über kausale Zusammenhänge zwischen Systemgüte, Erwartungshaltung und Benutzerzufriedenheit sind daher im Rahmen des ersten Experiments noch nicht möglich.

In der zweiten Studie wird aufbauend auf den Ergebnissen der ersten Untersuchung ein stärkeres Gewicht auf die Steuerbarkeit der Zufriedenheitsreaktion durch die Erwartungshaltung gelegt. Das zu diesem Zweck entwickelte Untersuchungsdesign stellt den Vergleich zweier Systeme in den Vordergrund und berücksichtigt in der Analyse der Suchergebniswahrnehmung acht verschiedene Untersuchungsgruppen: Insbesondere wird zwischen Testpersonen unterschieden,

die beide Suchaufgaben mit dem besseren, mit dem schlechteren oder je eine Aufgabe pro System bearbeiten. Da die Testpersonen im Zuge der Erwartungsmanipulation mitgeteilt bekommen, dass sie entweder die erste oder die zweite der beiden Suchaufgaben mit dem besseren oder dem schlechteren System bearbeiten, ergibt sich somit ein 2×4-Design, bei dessen Auswertung im Fall der zweiten Aufgabe neben dem Einfluss von Systemleistung und Erwartungshaltung zusätzlich die Qualität des zuerst genutzten Systems berücksichtigt werden muss. Studie 2 bietet damit zum einen die Möglichkeit, die in Studie 1 gewonnenen Ergebnisse zu überprüfen. Darüber hinaus trägt die verbesserte Erwartungsmanipulation dazu bei, dass diesmal auch die Erwartungshaltung als signifikante Einflussgröße auf die Suchergebniswahrnehmung bestätigt werden kann. Insbesondere zeigt sich, dass der Einfluss der Erwartungshaltung auf die Benutzerzufriedenheit einer zeitlichen Dynamik zu unterliegen scheint und die diesbezüglich signifikanten Unterschiede mit zunehmender Gewöhnung nachlassen. Das zur Verbesserung der Erwartungsmanipulation gewählte 2×4-Design erweist sich allerdings für die statistische Auswertung als ungünstig, wie die geringen Fallzahlen in SP_B sowie die zur Auswertung der zweiten Aufgabe notwendige zusätzliche Berücksichtigung des ersten Systems verdeutlichen. Vor allem bezogen auf die Befunde zur Zufriedenheitsdynamik ist aus den genannten Gründen nicht auszuschließen, dass es sich hier um eine Auswirkung des untersuchungsmethodischen Vorgehens handelt, weshalb die Analyse der dynamischen Abhängigkeit des Zufriedenheitsurteils zentraler Bestandteil der dritten Studie ist.

In Studie 3 schließlich sollen die Vorteile der ersten beiden Studien zusammengeführt werden. Dazu wird, um eine ähnliche Manipulationsstärke wie in Studie 2 zu erreichen, den Teilnehmern weiterhin der Vergleich zweier unterschiedlicher Suchmaschinen suggeriert. Um jedoch gleichzeitig den Komplexitätsgrad des Untersuchungsdesigns zu reduzieren, bearbeitet jede Testperson, ähnlich wie in Studie 1, alle drei Suchaufgaben mit derselben Suchmaschine. Bei der Entwicklung des Untersuchungsdesigns wird weiterhin Wert darauf gelegt, dass diesmal sowohl Selbstauskunfts- als auch Verhaltensmaße im zeitlichen Verlauf untersucht werden können. Deshalb wird im Zuge des dritten Experiments neben der Benutzerleistung auch die Benutzerzufriedenheit direkt im Anschluss an die Bearbeitung der einzelnen Suchaufgaben erfasst. Im Rahmen der zusätzlich durchgeführten Mittelwertanalysen trägt die Einbeziehung mehrerer Suchaufgaben überdies zu einer besseren Generalisierbarkeit der Ergebnisse bei. Ein weiteres Anliegen bei der Erarbeitung des Untersuchungsdesigns besteht darin, diesmal auch die Relevanzurteile der Testpersonen genauer analysieren zu können, weshalb eine feinere Skala zur Ausdifferenzierung der Relevanzurteile verwendet wird. Die Schwäche dieses Untersuchungsdesigns liegt in dem möglicherweise nicht gegebenen Vorwissen der Teilnehmer in Bezug auf das dritte Suchthema, in dem es um den Einsatz von Wikis im Schulunterricht geht (vgl. Abschn. 7.4.1). Da verhältnismäßig viele Teilnehmer in diesem Fall zunächst allgemeinere Suchbegriffe wie *wikis* oder *nutzung wikis* in das Suchfeld eingeben, kann vermutet werden, dass das Konzept eines Wikis in der betrachteten Stichprobe nicht hinreichend bekannt ist oder einzelne Teilnehmer die Aufgabenbeschreibung nicht gründlich lesen und infolgedessen Suchanfragen eingeben, die nur einen Teilaspekt des Suchthemas adressieren. Begegnet wird dieser Situation mit dem im Kontext der Einordnung des theoretischen und empirischen Beitrags dieser Arbeit (vgl. Abschn. 9.1.1) diskutierten stichprobendifferenzierenden Auswertungskonzept, durch das sichergestellt

wird, dass derartige Auffälligkeiten ausreichend Berücksichtigung finden.

Die Diskussion der drei experimentellen Studien verdeutlicht die Kompromisse, die bei einer validen Untersuchung des Informationssuchverhaltens erforderlich sind. In der gemeinsamen Betrachtung zeigt sich jedoch, dass sich die diskutierten Stärken und Schwächen der einzelnen Untersuchungsdesigns gegenseitig ausgleichen, sodass sich insgesamt ein einheitliches Zusammenhangsmuster ergibt, das zu einem tieferen Verständnis der untersuchten Wahrnehmungsphänomene beiträgt.

9.1.4. Zur Zuverlässigkeit der Ergebnisse dieser Arbeit

Der folgende Abschnitt diskutiert die Zuverlässigkeit der erhaltenen Ergebnisse. Da die inhaltliche Zuverlässigkeit bereits ausführlich in Kapitel 8 behandelt wird, richtet dieser Abschnitt den Blick auf das untersuchungsmethodische Vorgehen und die damit verbundene Absicherung der theoretischen Interpretation der Forschungsergebnisse. Diesbezüglich ist es wichtig, dass die zum Zeitpunkt der Untersuchungsplanung getroffenen Entscheidungen konsequent umgesetzt und überprüft werden. Konkret werden im Folgenden sechs Gütekriterien herausgearbeitet, die dabei helfen, die Zuverlässigkeit der Ergebnisse genauer einschätzen zu können.

Das erste Gütekriterium betrifft die Größe der untersuchten Stichproben. Ein sich aus dem gewählten Between-Subjects-Design ergebender Nachteil besteht in der höheren Anzahl von Testpersonen, die benötigt werden, um statistisch gesicherte Aussagen treffen zu können. Zwar werden die im Rahmen dieser Arbeit angestrebten Stichprobenumfänge von 20 Probanden pro Untersuchungsgruppe in den meisten Fällen erreicht, jedoch stellt sich insbesondere im Zuge der stichprobendifferenzierenden Auswertung heraus, dass durch den Ausschluss kritischer Fälle einige Analysen nicht durchgeführt werden können. Allerdings lässt sich argumentieren, dass die Zuverlässigkeit der Ergebnisse durch die Replikation der zentralen Befunde als ausreichend gesichert bezeichnet werden kann. Dies schließt sowohl die Replikation auf Ebene der einzelnen Studien als auch die Replikation auf Ebene der unterschiedlichen Datenqualitätsstufen mit ein. Auch das stichprobendifferenzierende Auswertungskonzept selbst kann als Gütekriterium zur Bestätigung der Zuverlässigkeit der Ergebnisse herangezogen werden. Dabei gilt: Lässt sich ein in $SP_{A,OT}$ nachgewiesener Befund durch weitere Überprüfungen bestätigen, nimmt seine Bedeutung zu. Dies ermöglicht eine optimale Nutzung der Daten, indem eine sehr konservative mit einer weniger restriktiven Fallauswahl verglichen und in einem einer Kreuzvalidierung ähnlichen Verfahren abgesichert wird. Hervorzuheben ist, dass in dieser Arbeit nicht nur die beiden Datensätze SP_A und SP_B sondern darüber hinaus im Zusammenhang mit der Überprüfung der Gütekriterien auch der Ausschluss einzelner auffällig gewordener Fallgruppen untersucht wird.

Da davon auszugehen ist, dass sowohl die Auswahl als auch die Bewertung von Informationsobjekten durch die Motivation der Benutzer mitbestimmt wird (vgl. Abschn. 2.2.1) und der Grad der Motivation von der empfundenen Selbstbestimmung abhängt (Ryan u. Deci, 2000), ist es in einem kontrollierten Experiment weiterhin wichtig, entsprechende motivationsfördernde Maßnahmen zu ergreifen, um bei den Probanden das Gefühl der Selbstbestimmung zu unterstützen. Dies wird in der vorliegenden Arbeit sowohl durch die Wahl der Suchthemen als auch die Möglichkeit, die Aufgaben eigenmächtig abzuschließen, berücksichtigt. Zwar ist, wie in Abschnitt 9.1.3 dargelegt, nicht auszuschließen, dass bspw. die Auffälligkeiten in Verbindung mit dem Wiki-Thema auf eine mangelnde Motivation der Testpersonen zurückzuführen sind, doch

lassen die Ergebnisse aller drei Studien insgesamt den Schluss zu, dass seitens der Probanden ein ausreichendes Maß an Interesse und Motivation besteht. Diese Annahme wird durch allgemeine positive Signale im Anschluss an die Testdurchführung sowie im Rahmen der Beantwortung der offenen Frage untermauert. Darüber hinaus führt die im zweiten Experiment getestete Einbeziehung der Ausgangsmotivation als Kovariate weder für die Benutzerleistung noch für die Benutzerzufriedenheit zu grundsätzlich anderen als den bereits dargestellten Ergebnissen.

Wie eingangs bereits erwähnt, ist auch eine einheitliche und zutreffende Ermittlung der abhängigen Variablen für die Zuverlässigkeit der Ergebnisse von großer Bedeutung. Als zentrales Konzept für die Beurteilung des Sucherfolgs gilt in der IR-Evaluation die Relevanz der Dokumente. Die im Zuge der Erstellung der Testkorpora erfolgte Absicherung der Relevanzurteile durch mindestens zwei unabhängige Juroren stellt in diesem Zusammenhang ein weiteres Gütekriterium dar. Die zur Überprüfung der Konsistenz der entwickelten Testkorpora bestimmten Interrater-Reliabilitäten bewegen sich in beiden Studien (Exp. 2 u. 3) auf einem relativ moderaten Niveau. Unter der Annahme, dass Dokumente umso eindeutiger in eine bestimmte Relevanzkategorie fallen, je weniger Jurorenurteile benötigt werden, wird zudem eine Obergrenze von drei Relevanzurteilen zur Aufnahme in das finale Korpus festgelegt. Im Fall der Zufriedenheitsermittlung kann die Verwendung eines bereits etablierten Fragebogeninstruments wie auch die Vorgehensweise zur Skalenbildung als Gütekriterium gelten. Zwar kann die Faktorenstruktur des EUCS-Instruments nur in Teilen repliziert werden, jedoch lassen sich die darüber hinaus faktoranalytisch ermittelten Skalen inhaltlich gut interpretieren und sind ausreichend reliabel. Des Weiteren erweisen sich die zur Beurteilung der einzelnen Skalen herangezogenen Gütekriterien über verschiedene Suchthemen und Fallgruppen hinweg als hinreichend stabil.

Der letzte Punkt betrifft die statistische Auswertung der Ergebnisse. Besonders hervorzuheben ist hier die konservative Handhabung der Datenanalyse. So wird im Fall von Topiceffekten eine nach Versuchsgruppe und Suchthema balancierte Fallauswahl vorgenommen, um die Zuverlässigkeit der Ergebnisse zu gewährleisten. Zusätzlich wird der konservative Charakter der im Rahmen dieser Arbeit durchgeführten Analyse durch die im Fall von Voraussetzungsverletzungen erfolgte Anwendung robuster Verfahren verstärkt. Auch der Einsatz von Kovarianzanalysen zur Erhöhung der internen Validität der Daten ist letztlich in diesem Licht zu sehen.

Zusammenfassend macht auch die Diskussion des untersuchungsmethodischen Vorgehens deutlich, dass ein zentraler wissenschaftlicher Beitrag dieser Arbeit in der systematischen und breit angelegten Untersuchung des Zusammenspiels von Systemqualität und Erwartungshaltung liegt. Durch die Kombination mehrerer Methoden der Datenerhebung und -auswertung lässt sich die Bedeutung der Erkenntnisse und Befunde festigen. Gleichzeitig ergeben sich weitere Perspektiven auf den Untersuchungsgegenstand, die es gestatten diesen möglichst umfassend zu betrachten.

9.2. Fazit und Ausblick

Bevor die Erkenntnisse dieser Arbeit in einem abschließenden Fazit zusammengefasst werden, stellen die folgenden beiden Abschnitte einige Überlegungen und Möglichkeiten zur Weiterführung der behandelten Forschungsfragen sowie zur praktischen Nutzung der Ergebnisse dar.

9.2.1. Überlegungen und Möglichkeiten zur weiterführenden Forschung

Im Zuge der Interpretation der Ergebnisse wird deutlich, dass die vorliegende Arbeit zwar einen bedeutenden Beitrag zur Aufklärung der Rolle von Systemgüte und Erwartungshaltung für einen erfolgreichen Suchprozess leisten kann, sich jedoch gleichzeitig weiterführender Forschungsbedarf ergibt. Neben den bereits im Rahmen der vorangegangenen Diskussion dargestellten Aspekten, stellen insbesondere die im Folgenden dargelegten Forschungsrichtungen interessante und wünschenswerte Anknüpfungspunkte dar.

Fragebogeninventar zur mehrdimensionalen Erfassung der Suchzufriedenheit Angesichts der wachsenden Zahl interaktiver IR-Studien wäre es wünschenswert, die Vergleichbarkeit zwischen den erzielten Ergebnissen zu verbessern. Ein wichtiger Schritt in diese Richtung wäre die Erstellung und Validierung eines mehrsprachigen, flexiblen und umfassenden Fragebogeninventars, das unterschiedliche Aspekte des wahrgenommenen Sucherfolgs und der Benutzerzufriedenheit adressiert. Mit dem in dieser Arbeit verwendeten EUCS-Instrument existiert ein solcher Fragebogen bereits im Kontext der Informationssystemforschung. Ein speziell an die Bedürfnisse interaktiver IR-Evaluationen angepasstes Instrument fehlt jedoch bisher. Aufbauend auf den Erfahrungen dieser Arbeit und der zitierten Literatur könnte ein solches Fragebogeninstrument entwickelt werden, das auch auf die jüngsten Interessen in diesem Bereich – wie z.B. Selbstwirksamkeit (Tsai u. Tsai, 2003; Monoi et al., 2005; Chiou u. Wan, 2007) und Zeitwahrnehmung (Luo et al., 2017) – eingeht. Idealerweise sollte ein solcher Fragebogen der IR-Community als gebrauchsfertige Komponente zur Durchführung von IIR-Studien zur Verfügung gestellt werden, bspw. als limesurvey-Fragebogen mit den entsprechenden R-Skripten für die anschließende Auswertung.

Untersuchung unterschiedlicher Dimensionen und der Dynamik von Benutzererwartungen Im Kontext dieser Arbeit werden Benutzererwartungen manipuliert, indem Probanden mitgeteilt wird, welche Systemleistung sie erwarten können. Unter Berücksichtigung der Benutzerbeteiligung am Suchprozess, wäre es interessant, über diese spezifische Operationalisierung hinauszugehen und andere Priming-Effekte zu untersuchen. So wäre es bspw. möglich, die Teilnehmer die entsprechende Erwartungshaltung während einer Trainingsphase selbst entwickeln zu lassen. In diesem Zusammenhang bieten sich unterschiedliche Manipulationsstrategien an, die jeweils andere Aspekte der Nutzererwartung berühren und auf diese Weise weiter erforscht werden könnten: Änderung der Systemleistung nach der ersten Trainingsphase, Änderung des Schwierigkeitsgrads der Suchaufgaben während der Suchsitzung oder Rückmeldung bezüglich der Suchleistung (z.B. welcher Prozentsatz der relevanten Dokumente gefunden wurde). Insbesondere wäre es interessant zu beobachten, ob diese erfahrungsorientierten Erwartungen eine stärkere und länger anhaltende Wirkung haben als kommunizierte Erwartungen, wie sie im Rahmen dieser Arbeit verwendet werden. Dieser letzte Punkt berührt auch die allgemeinere Frage, wie sich die Erwartungen der Nutzer im Laufe der Zeit entwickeln. Die Ergebnisse des dritten Experiments geben Anlass zu der Vermutung, dass niedrige Erwartungen sich in einer Abwärtsspirale verstärken können (vgl. Abschn. 8.3.2): Niedrige Anfangserwartungen führen zur Wahrnehmung einer niedrigen Systemqualität, was wiederum eine Absenkung der Erwartungen in Bezug auf die nächste Suchaufgabe zur Folge hat. Daher wird eine solche Rückkopplungsschleife zu einer stetig abnehmenden Benutzerzufriedenheit führen. Allerdings scheinen Anfangserwartungen in anderen

Nutzungssituationen ihren Einfluss mit der Zeit zu verlieren und allmählich durch die aktuelle Systemqualität ersetzt zu werden, während in wieder anderen Fällen eine Wiederbelebung der initialen Erwartungshaltung nach einer Reihe von Suchaufgaben zu beobachten ist. Dies zeigt, dass die Dynamik von Benutzererwartungen ein sehr reichhaltiges und kontextabhängiges Phänomen darstellt. Eine lohnenswerte Forschungsrichtung stellt daher die tiefergehende Analyse der Frage dar, welche Dimensionen der Benutzerzufriedenheit zu welchem Verhalten führen. Darüber hinaus ließe sich das Zusammenspiel von Erwartungen und Selbstwirksamkeit im Kontext des Suchprozesses genauer untersuchen, so z.B. die Frage, ob Änderungen der Selbstwirksamkeitsüberzeugung in Folge erfolgreicher oder erfolgloser Suchen mithilfe eines gegebenen Systems durch die initiale Erwartungshaltung moderiert werden.

Fokussierung auf die Dimensionalität der Relevanzwahrnehmung Auch wenn die Anpassung von Relevanzkriterien aufgrund der Exposition mit unterschiedlichen Systemqualitätsstufen im Rahmen dieser Arbeit experimentell verifiziert werden kann, ergeben sich doch auch eine Reihe weiterführender Fragestellungen. So geben die Ergebnisse dieser Arbeit bspw. keinen Aufschluss darüber, wie diese Wahrnehmungstendenz quantitativ mit der Systemqualität zusammenhängt. Zum einen wäre es möglich, dass es sich um eine stetige Änderung handelt, die mit steigender Systemqualität zu immer strengeren Relevanzkriterien führt, während andererseits auch objektive oder individuelle Schwellenwerte existieren könnten, die die Anwendung restriktiverer Relevanzurteile auslösen. Ebenso wäre es theoretisch denkbar, dass sich der beobachtete Anpassungseffekt von einem gewissen Systemgütenniveau an wieder umkehrt. Von praktischer Relevanz ist dabei vor dem Hintergrund all dieser Punkte die Frage, ob eine optimale Systemqualität oder ein Systemleistungsbereich bestimmbar sind, die zu einer maximalen Benutzerzufriedenheit führen. Um diese Fragen untersuchen zu können, sind Benutzerstudien mit feineren Variationen und einer größeren Anzahl unterschiedlicher Systemqualitäten notwendig. Darüber hinaus wäre es interessant, die Messung der wahrgenommenen Relevanz zu verfeinern: Neben der allgemeinen Relevanz eines Dokuments im Hinblick auf die zu bearbeitende Suchaufgabe könnte, ähnlich wie bei Jiang et al. (2017), der hinzukommende Informationsgewinn eines Dokuments zum Auffindezeitpunkt, seine Glaubwürdigkeit sowie der individuell wahrgenommene Aufwand zum Auffinden bewertet werden.

Übertragung der Fragestellung auf die mobile Informationssuche Die Allgegenwärtigkeit mobiler Endgeräte erzeugt neue Kontexte für die Informationssuche. Dies geht mit einer Änderung des Wie, Wo und Was der Suchsituation einher. Intelligente Assistenten wie Google Now, Siri, Cortana und Alexa eröffnen neue Wege der Interaktion mit Suchmaschinen und sowohl Relevanzurteile als auch Suchaufgaben werden in noch stärkerem Maße abhängig von geografischen und situativen Gegebenheiten. Die Forschung in Bezug auf die Auswirkungen dieser neuen und aufregenden Möglichkeiten beginnt sich gerade erst als eigene Forschungsrichtung zu etablieren. Um zu diesem Forschungsfeld aus einer informationswissenschaftlichen Perspektive beizutragen, bieten sich bspw. vergleichende Benutzerstudien an, um Unterschiede im Benutzerverhalten zwischen der Interaktion mit intelligenten Assistenten und klassischen Suchmaschinen zu identifizieren. Mit Blick auf die Erwartungshaltung der Benutzer ist anzunehmen, dass die durch die Gegenwart eines intelligenten Assistenten entstandene Kommunikationssituation die Wahrnehmung der Suchergebnisse zusätzlich beeinflusst. Neben der Überprüfung der im Rahmen

der Desktop-Suche gewonnenen Befunde dieser Arbeit, erscheinen insbesondere auch folgende Forschungsfragen interessant: Ändert sich die Struktur der gestellten Suchanfragen durch die wiederholte Interaktion mit einem Assistenten, z.B. hinsichtlich der Verwendung von natürlicher Sprache? Gibt es Unterschiede in der Zuschreibung von Sucherfolg und -misserfolg? Verändert die Gegenwart eines intelligenten Assistenten die Glaubwürdigkeit der Suchergebnisse, d.h. ihre wahrgenommene Relevanz?

Zusammenfassend bietet dieser Abschnitt einen ersten Überblick über mögliche weiterführende Forschungsbedarfe. Die aufgezeigten Anknüpfungspunkte lassen erkennen, dass es zielführend erscheint, zukünftige Forschungsaktivitäten auf zwei Themenbereiche zu fokussieren. Zum einen ergeben sich konkrete methodische Problemstellungen (wie z.B. die Entwicklung eines Fragebogeninventars zur Erfassung IR-spezifischer Zufriedenheitsdimensionen), zum anderen stellt die Dimensionalität der betrachteten Größen einen wichtigen Ansatzpunkt für weitere Forschungsarbeiten dar. Darüber hinaus ergeben sich aus einer Übertragung der betrachteten Fragestellungen auf erweiterte Anwendungskontexte (wie z.B. die mobile Suche und intelligente Assistenten) Bezüge zu weiteren aktuellen Forschungsrichtungen.

9.2.2. Überlegungen zur praktischen Relevanz der Befunde

Vor dem Hintergrund der durchgeführten Nutzerstudien kann gezeigt werden, dass Benutzererwartungen auf vielfältige Weise in den Suchprozess hineinwirken. Dies betrifft zum einen den Einfluss der bereits angesprochenen Anpassungseffekte auf die Relevanzwahrnehmung und ihre Auswirkungen auf die Benutzerzufriedenheit. Zum anderen zeigt sich jedoch auch, dass Standardmodelle der Kundenzufriedenheitsforschung, wie bspw. das C/D-Paradigma, nur bedingt auf den Kontext des Suchprozesses übertragbar sind. Dieser Abschnitt diskutiert praktische Implikationen, die sich aus diesen Untersuchungsergebnissen ergeben. Für die Entwicklung bzw. den Betrieb effektiver Suchsysteme lassen sich im Wesentlichen drei Ansatzpunkte identifizieren, die die Auslieferung individuell angepasster Suchergebnisse, das Erwartungsmanagement sowie die Sicherstellung einer hohen Rankingqualität betreffen.

Der im Rahmen dieser Arbeit gewonnene Einblick in das komplexe Zusammenspiel von Systemqualität und Benutzererwartung zeigt zunächst, dass die Unterschiede zwischen den verglichenen Systemqualitäten im Fall der Benutzerzufriedenheit weitgehend verschwinden. Der Erfolg einer Suchmaschine baut in diesem Sinne somit weniger auf objektiv gegebenen, sondern vom Benutzer subjektiv wahrgenommenen Wettbewerbsvorteilen auf. Die Erfüllung der Benutzerbedürfnisse sollte daher für Suchmaschinenbetreiber eine hohe Priorität haben. Schnell wird klar, dass es mit zunehmend vielfältiger Nutzerschaft immer schwieriger wird, allen Wünschen und Erwartungen mit einem statischen Angebot gerecht zu werden. Eine mögliche Strategie, um diesen Anforderungen zu entsprechen, stellt bspw. die Auslieferung individueller Suchergebnisse dar, für deren Bereitstellung der Suchdienst auf vergangene Suchaktivitäten zurückgreift und diese mit weiteren Hintergrundinformationen, wie bspw. der geographischen Position des Suchenden, verknüpft. Mit personalisierten Suchergebnissen wird dieser Ansatz bspw. von Google und Bing schon seit längerem verfolgt. Allerdings zeigt sich, dass derartige Maßnahmen – gerade im Zusammenhang mit Erwartungen – auch dem Ziel, das Sucherlebnis zu verbessern, entgegen wirken können. Zum einen kann die unkontrollierte Verknüpfung von persönlichen Daten Widerstand und Misstrauen auslösen, wenn Benutzer ihre Privatsphäre verletzt sehen und

Datenmissbrauch befürchten. Zum anderen besteht die Gefahr, dass auf diese Weise die Interessensvielfalt der Benutzer nicht in ausreichendem Maße berücksichtigt wird und die präsentierten Ergebnisse nunmehr lediglich eine Spiegelung des bisherigen Klickverhaltens darstellen, wodurch eine breite Informationssuche, die ein Thema von unterschiedlichen Perspektiven beleuchten soll, möglicherweise erschwert wird. Damit gerät das externe Marketing der Suchmaschinen in den Fokus: Bietet ein Suchdienst seiner Nutzerschaft zusätzliche Optionen wie personalisierte Suchergebnisse an, ist eine transparente Darstellung der verwendeten Datengrundlage bzw. der erhobenen Nutzerdaten erforderlich, die dem Nutzer gleichzeitig ein ausreichendes Maß an Kontrolle über den gewünschten Grad der Personalisierung ermöglicht.

Vor dem Hintergrund der in dieser Arbeit gewonnenen Erkenntnisse scheint darüber hinaus ein positives Erwartungsmanagement wichtig. Insbesondere die Beobachtung, dass erwartungsbezogene Hinweisreize vorwiegend in Situationen herangezogen werden, in welchen das eigene Vorwissen zur Beurteilung und Verarbeitung von neuen Informationen nicht ausreicht, macht dies deutlich. Auch der im zweiten Experiment auftretende Placebo-Effekt in Bezug auf die wahrgenommene Benutzerfreundlichkeit der beiden sich unter Usability-Gesichtspunkten de facto nur hinsichtlich der Farbe der Benutzeroberfläche unterscheidenden Testsysteme ist in dieser Hinsicht zu deuten. Phänomenologisch interessant scheint in diesem Zusammenhang, dass qualitätsbezogene Erwartungen als Beurteilungsunterstützung herangezogen werden können. Mit anderen Worten bedeutet dies, dass Nutzer, die bereits im Vorfeld der Systemnutzung eine hohe Meinung von der Systemqualität haben, die gefundenen Informationsobjekte und in der Folge auch den Erfolg der gesamten Suche höher bewerten. In gewisser Weise werden so die Ergebnisse von Joachims et al. (2005), Pan et al. (2007) und Keane et al. (2008) bestätigt, wonach sich eine Wiederholungserwartung einstellt, wenn Nutzer sehen, dass Suchergebnisse meistens in der Reihenfolge ihrer Relevanz angezeigt werden. Um dieses Vertrauen nicht zu enttäuschen und die Zufriedenheit der Benutzer sicherzustellen, ist es deshalb zum einen notwendig, eine gleichbleibende Qualität der Suchfunktionalitäten anzubieten. Zum anderen legen diese Befunde nahe, dass eine positive Imagearbeit und eine überzeugende Kommunikation nach außen die Benutzerzufriedenheit positiv beeinflussen können, da beide Faktoren dazu dienen, die Meinungsbildung der Nutzer mitzubestimmen. Inwiefern sich darüber hinaus eine gute Usability und ein sympathisches Design positiv auf die Benutzerzufriedenheit auswirken, kann hingegen ausgehend von den Ergebnissen dieser Arbeit nicht beantwortet werden. Es ist jedoch anzunehmen, dass auch sie die Erwartungsbildung der Benutzer und damit ihr Zufriedenheitsurteil in ähnlicher Weise vorprägen können.

Die im Rahmen dieser Arbeit nachgewiesenen Anpassungsreaktionen, geben Hinweise auf weitere Stellschrauben zur Erhöhung des wahrgenommenen Sucherfolgs. Die Tatsache, dass der systembedingte Anpassungseffekt der Relevanzwahrnehmung den Einfluss der Systemleistung auf die Benutzerzufriedenheit zu überlagern scheint, gibt hier Anlass zu der Vermutung, dass Maßnahmen zur Verhinderung dieses Verhaltens zu einer Verbesserung der Zufriedenheit beitragen können. In diesem Zusammenhang erweist sich insbesondere die Beobachtung, dass die von Scholer und Turpin (2008) und Scholer et al. (2008) eingeführten individuellen Relevanzschwellen zusätzlich eine kontextabhängige Komponente aufweisen als lohnenswerter Ansatzpunkt. Konkret zeigt sich, dass derartige Anpassungsreaktionen vorzugsweise im Kontext von Doku-

menten mit geringerer Relevanz auftreten, sodass weiterhin davon auszugehen ist, dass eine Verbesserung der Rankingqualität einen positiven Einfluss auf die Benutzerzufriedenheit haben sollte.

Ausgehend von den Ergebnissen dieser Arbeit lassen sich somit unmittelbar drei Handlungsempfehlungen zur Verbesserung des Nutzererlebnisses bzw. der Nutzerzufriedenheit identifizieren. Während die beiden zuletzt diskutierten Maßnahmen konkret die Wirkungsweise von Systemgüte und Erwartungshaltung berücksichtigen, adressiert der erste Ansatz allgemeiner die Verwendung personalisierter Suchergebnisse aus der Perspektive der Benutzererwartung. Darüber hinaus machen die Ergebnisse dieser Arbeit deutlich, dass eine direkte Übertragung von Erkenntnissen aus der Kundenzufriedenheitsforschung auf den Suchkontext nicht ohne weiteres möglich ist, da die aktive Beteiligung des Nutzers am Suchergebnis in stärkerem Maße berücksichtigt werden muss.

9.2.3. Fazit

Betrachtet man das breite Spektrum aktueller IR-Ansätze zur Evaluation von Suchsystemen aus einer historischen Perspektive, so rückt der Benutzer als aktiv am Ergebnis der Suche beteiligter Einflussfaktor zunehmend in den Mittelpunkt aktueller Forschung. Während im Rahmen des Cranfield-Paradigmas noch mehrheitlich von einem systemzentrierten Relevanzbegriff ausgegangen wird, der allein auf einem Vergleich von Suchanfrage und Suchergebnis beruht, wird nun zunehmend auch der situative Kontext, in dem die einzelnen Informationsobjekte durch den Nutzer wahrgenommen und verarbeitet werden, berücksichtigt. Allerdings lässt sich diese situative Relevanz weniger präzise erfassen, da sich der kontextspezifische Nutzen eines konkreten Dokuments aus der Betrachtung einer Vielzahl verschiedener Aspekte wie Aufgabenschwierigkeit, Vorwissen oder der Reihenfolge der präsentierten Informationsobjekte ergibt, die nur indirekt im Zusammenhang mit der eigentlichen Suchanfrage stehen. Die Integration des Benutzers in den Evaluierungsprozess geht also in natürlicher Weise mit einer Erweiterung des Relevanzbegriffs einher, was jedoch gleichzeitig eine Zunahme des Komplexitätsgrad in Bezug auf die experimentelle Umsetzung und Analyse nach sich zieht. Dennoch stellt die Möglichkeit einer ganzheitlichen Betrachtung der Relevanzbeurteilung eine wesentliche Stärke benutzerorientierter Evaluierungsansätze dar. Die Ergebnisse dieser Arbeit unterstreichen diese situative Komponente des Relevanzbegriffs anhand der beschriebenen Anpassungseffekte: Benutzer passen ihre Relevanzkriterien flexibel an die jeweilige Suchsituation an, wobei in der vorliegenden Arbeit sowohl die präsentierte Systemgüte als auch die erhaltene Erwartungshaltung zu solch einer Anpassungsreaktion führen können. Darüber hinaus wird deutlich, dass dieses Verhalten im Wesentlichen auf den Grenzbereich zwischen relevanten und irrelevanten Dokumenten beschränkt bleibt und somit als eine Reaktion auf einen Zustand der Unsicherheit verstanden werden kann, der zu einer Verschiebung des individuellen Relevanzschwellenwerts führt. Speziell die Untersuchung von Erwartungseinflüssen gewinnt damit für die IR-Forschung an Bedeutung: Erwartungen werden nicht mehr nur als Ausdruck von individuellen Bedürfnissen oder als Folge eigener Erfahrungen angesehen, sondern als Träger relevanter Information für die wechselseitige Interpretation suchbezogener Handlungen begriffen. Die in dieser Arbeit untersuchte Übertragbarkeit des C/D-Paradigmas auf den Kontext der Informationssuche macht darüber hinaus deutlich, dass das Zusammenspiel von Systemgüte, Benutzererwartungen und Benutzerzufriedenheit, vermutlich

aufgrund der stärkeren Beteiligung des Nutzers am Suchprozess, nicht notwendigerweise den aus der Kundenzufriedenheit bekannten Wirkungsmechanismen folgt. Die Analysen zeigen jedoch, dass grundsätzlich ein positiver Zusammenhang zwischen den untersuchten unabhängigen Variablen Erwartungshaltung und Systemleistung und der Benutzerzufriedenheit zu bestehen scheint.

Die Befunde dieser Arbeit unterstützen somit den gegenwärtigen Trend zur benutzerorientierten Untersuchung des IR-Prozesses sowie zur Fokussierung auf verhaltensbezogene und kognitive Bewältigungsstrategien als zentrale Determinanten des Sucherfolgs. Gleichzeitig machen sie deutlich, dass individuelle Unterschiede zwischen den Benutzern nicht vernachlässigt werden sollten. Die Ergebnisse verdeutlichen darüber hinaus, dass eine gemeinsame Betrachtung von Systemqualität und Erwartungshaltung die einem erfolgreichen Suchprozess zugrundeliegenden Einflussfaktoren und deren Wirkung auf die Wahrnehmung des individuellen Sucherfolgs besser abbilden kann als ihre getrennte Untersuchung. Dies unterstreichen insbesondere die nachgewiesenen Wechselwirkungen zwischen Systemgüte und Erwartungshaltung. Darüber hinaus wird durch die gemeinsame Untersuchung von objektiven Benutzerleistungsmaßen und subjektiven Zufriedenheitsindikatoren deutlich, dass die wahrgenommene Systemqualität stärker als das objektiv vorhandene Leistungsniveau Eingang in das finale Zufrienheitsurteil findet. Der Erfolg und der Grad der Zufriedenheit mit einer Suche sind somit ganz im Sinne des nutzerzentrierten Evaluierungsparadigmas nicht allein durch die objektive Systemqualität determiniert, sondern beruhen darüber hinaus auch auf individuellen Einflussfaktoren, wie eben der Erwartungshaltung des Nutzers.

Literaturverzeichnis

Al-Maskari et al. 2006
Al-Maskari, A.; Clough, P. u. Sanderson, M. (2006): „Users' effectiveness and satisfaction for image retrieval". In: *Proc. LWA: Workshop on Information Retrieval of the Special Interest Group Information Retrieval* (Hildesheim, 9.–11. Okt. 2006). Hildesheim: Universität Hildesheim, S. 84–88 (siehe S. 3, 43, 63, 80, 86 f., 98 f., 101, 111, 114, 146, 323).

Al-Maskari u. Sanderson 2006
Al-Maskari, A. u. Sanderson, M. (2006): „The effects of topic familiarity on user search behavior in question answering systems". In: *Proc. LWA: Workshop on Information Retrieval of the Special Interest Group Information Retrieval* (Hildesheim, 9.–11. Okt. 2006). Hildesheim: Universität Hildesheim, S. 132–137 (siehe S. 30, 125).

Al-Maskari u. Sanderson 2010
Al-Maskari, A. u. Sanderson, M. (2010): „A review of factors influencing user satisfaction in information retrieval". In: *J. Am. Soc. Inf. Sci. Technol.* 61 (5), S. 859–868 (siehe S. 75 f., 78, 80, 86 f., 114, 116, 233, 335).

Al-Maskari et al. 2007
Al-Maskari, A.; Sanderson, M. u. Clough, P. (2007): „The relationship between IR effectiveness measures and user satisfaction". In: *Proc. 30th SIGIR* (Amsterdam, 23.–27. Juli 2007). New York: ACM, S. 773–774 (siehe S. 77, 86, 97, 118, 162).

Al-Maskari et al. 2008a
Al-Maskari, A.; Sanderson, M. u. Clough, P. (2008a): „Relevance judgments between TREC and non-TREC assessors". In: *Proc. 31st SIGIR* (Singapore, 20.–24. Juli 2008). New York: ACM, S. 683–684 (siehe S. 58, 61, 85, 156, 195).

Al-Maskari et al. 2008b
Al-Maskari, A.; Sanderson, M.; Clough, P. u. Airio, E. (2008b): „The good and the bad system: Does the test collection predict users' effectiveness?" In: *Proc. 31st SIGIR* (Singapore, 20.–24. Juli 2008). New York: ACM, S. 59–66 (siehe S. 57 ff., 85, 114, 116, 156, 195, 226, 256, 321, 339).

Allan et al. 2005
Allan, J.; Carterette, B. u. Lewis, J. (2005): „When will information retrieval be "good enough"?" In: *Proc. 28th SIGIR* (Salvador, 15.–19. Aug. 2005). New York: ACM, S. 433–440 (siehe S. 43, 57 ff., 63, 85 f., 93 f., 99, 101, 107, 111, 114, 116, 195, 256, 321, 335, 339).

Anderson 2005
Anderson, T. D. (2005): „Relevance as process: Judgements in the context of scholarly research". In: *Inform. Res.* 10 (2) (siehe S. 48 f.).

Applegate 1993
Applegate, R. (1993): „Models of user satisfaction: Understanding false positives". In: *RQ* 32 (4), S. 525–539 (siehe S. 118).

Aula 2003
Aula, A. (2003): „Query formulation in web information search". In: *Proc. ICWI* (Algarve, 5.–8. Nov. 2003). IADIS, S. 403–410 (siehe S. 99).

Bailey u. Pearson 1983
Bailey, J. E. u. Pearson, S. W. (1983): „Development of a tool for measuring and analyzing computer user satisfaction". In: *Manag. Sci.* 29 (5), S. 530–545 (siehe S. 118 f.).

Bandura 1986
Bandura, A. (1986): *Social foundations of thought and action*. Englewood Cliffs: Prentice-Hall (siehe S. 16, 35 f.).

Bandura 2006
Bandura, A. (2006): „Guide for constructing self-efficacy scales". In: Pajares, F. u. Urdan, T. C., Pajares, F. u. Urdan, T. C. (Hrsg.): *Self-efficacy beliefs of adolescents*. Greenwich: IAP, S. 307–337 (siehe S. 37).

Barry 1994
Barry, C. L. (1994): „User-defined relevance criteria: An exploratory study“. In: *J. Am. Soc. Inf. Sci.* 45 (3), S. 149–159 (siehe S. 4, 50, 85).

Barry u. Schamber 1998
Barry, C. L. u. Schamber, L. (1998): „Users‘ criteria for relevance evaluation: A cross-situational comparison“. In: *Inf. Process. Manage.* 34 (2-3), S. 219–236 (siehe S. 50, 85).

Beierlein et al. 2013
Beierlein, C.; Kemper, C. J.; Kovaleva, A. u. Rammstedt, B. (2013): „Kurzskala zur Erfassung allgemeiner Selbstwirksamkeitserwartungen (ASKU)“. In: *Methoden Daten Anal.* 7 (2), S. 251–278 (siehe S. 16, 35–38).

Beitzel et al. 2004
Beitzel, S. M.; Jensen, E. C.; Chowdhury, A.; Grossman, D. u. Frieder, O. (2004): „Hourly analysis of a very large topically categorized web query log“. In: *Proc. 27th SIGIR* (Sheffield, 25.–29. Juli 2004). New York: ACM, S. 321–328 (siehe S. 90).

Belkin et al. 1999
Belkin, N.; Cool, C.; Head, J.; Jeng, J.; Kelly, D.; Lin, S.; Lobash, L.; Park, S.; Knepshield, S. P. u. Sikora, C. (1999): „Relevance feedback versus local context analysis as term suggestion devices: Rutgers' TREC-8 interactive track experience“. In: *Proc. 8th TREC* (Gaithersburg, 16.–19. Nov. 1999). NIST, S. 565–574 (siehe S. 101, 111).

Borlund 2000
Borlund, P. (2000): „Experimental components for the evaluation of interactive information retrieval systems“. In: *J. Doc.* 56 (1), S. 71–90 (siehe S. 49).

Borlund 2003a
Borlund, P. (2003a): „The concept of relevance in IR“. In: *J. Am. Soc. Inf. Sci. Technol.* 54 (10), S. 913–925 (siehe S. 48 f., 326).

Borlund 2003b
Borlund, P. (2003b): „The IIR evaluation model: A framework for evaluation of interactive information retrieval systems“. In: *Inform. Res.* 8 (3) (siehe S. 6).

Borlund u. Ingwersen 1997
Borlund, P. u. Ingwersen, P. (1997): „The development of a method for the evaluation of interactive information retrieval systems“. In: *J. Doc.* 53 (3), S. 225–250 (siehe S. 34, 49, 72, 99, 329).

Bortz 2005
Bortz, J. (2005): *Statistik: Für Human- und Sozialwissenschaftler*. 6. Aufl. Heidelberg: Springer (siehe S. 127 ff., 131–138, 273).

Bortz u. Döring 2006
Bortz, J. u. Döring, N. (2006): *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler*. 4. Aufl. Heidelberg: Springer (siehe S. 55, 103, 112, 117, 121, 123 f., 238 f.).

Boulding et al. 1993
Boulding, W.; Kalra, A.; Staelin, R. u. Zeithaml, V. A. (1993): „A dynamic process model of service quality: From expectations to behavioral intentions“. In: *J. Mark. Res.* 30 (1), S. 7 (siehe S. 70, 210).

Braschler 2002
Braschler, M. (2002): „CLEF 2001 - Overview of results“. In: *Proc. 2nd CLEF* (Darmstadt, 3.–4. Sep. 2001). LNCS 2406. Berlin: Springer, S. 9–26 (siehe S. 147).

Braschler 2004
Braschler, M. (2004): „CLEF 2003 - Overview of results“. In: *Proc. 4th CLEF* (Trondheim, 21.–22. Aug. 2003). Bd. 3237. Lecture Notes in Computer Science 3237. Berlin: Springer, S. 44–63 (siehe S. 147).

Bruce 1994
Bruce, H. W. (1994): „A cognitive view of the situational dynamism of user-centered relevance estimation“. In: *J. Am. Soc. Inf. Sci.* 45 (5), S. 142–148 (siehe S. 54, 82, 85, 94, 112 f.).

Buckley u. Voorhees 2004
Buckley, C. u. Voorhees, E. M. (2004): „Retrieval evaluation with incomplete information“. In: *Proc. 27th SIGIR* (Sheffield, 25.–29. Juli 2004). New York: ACM, S. 25–32 (siehe S. 106).

Buckley u. Voorhees 2005
Buckley, C. u. Voorhees, E. M. (2005): „Retrieval system evaluation“. In: Voorhees, E. M. u. Harman,

D. K., Voorhees, E. M. u. Harman, D. K. (Hrsg.): *TREC: Experiment and evaluation in information retrieval*. Cambridge: MIT Press, S. 53–75 (siehe S. 2).

Bunse 2000
Bunse, T. (2000): „Kundenzufriedenheit und Wartezeiten: Eine empirische Analyse für den Luftverkehr". Diss. Freiburg: Albert-Ludwigs-Universität (siehe S. 9, 15, 17, 66, 68).

Büttcher et al. 2010
Büttcher, S.; Clarke, C. L. A. u. Cormack, G. V. (2010): *Information retrieval: Implementing and evaluating search engines*. Cambridge: MIT Press (siehe S. 75).

Cadotte et al. 1978
Cadotte, E. R.; Woodruff, R. B. u. Jenkins, R. L. (1978): „Expectations and norms in models of consumer satisfaction". In: *J. Mark. Res.* 24 (3), S. 305–314 (siehe S. 65).

Cazan et al. 2016
Cazan, A.-M.; Cocorada, E. u. Maican, C. I. (2016): „Computer anxiety and attitudes towards the computer and the internet with romanian high-school and university students". In: *Comput. Human Behav.* 55, S. 258–267 (siehe S. 36).

Chadwick-Dias et al. 2003
Chadwick-Dias, A.; McNulty, M. u. Tullis, T. (2003): „Web usability and age: How design changes can improve performance". In: *Proc. CUU* (Vancouver, 10.–11. Nov. 2003). New York: ACM, S. 30–37 (siehe S. 40).

Chang et al. 2016
Chang, L.; Zhang, X. u. Huang, W. (2016): „The exploration of objective task difficulty and domain knowledge effects on users' query formulation". In: *Proc. 79th ASIST* (Kopenhagen, 14.–18. Okt. 2016). Silver Springs: ASIST (siehe S. 44).

Chen 1986
Chen, M. (1986): „Gender and computers: The beneficial effects of experience on attitudes". In: *J. Educ. Comput. Res.* 2 (3), S. 265–282 (siehe S. 36).

Chevalier et al. 2015
Chevalier, A.; Dommes, A. u. Marquié, J.-C. (2015): „Strategy and accuracy during information search on the web: Effects of age and complexity of the search questions". In: *Comput. Human Behav.* 53, S. 305–315 (siehe S. 38 f., 45, 91).

Chin u. Fu 2010
Chin, J. u. Fu, W.-T. (2010): „Interactive effects of age and interface differences on search strategies and performance". In: *Proc. 28th CHI* (Atlanta, 10.–15. Apr. 2010). New York: ACM, S. 403–412 (siehe S. 35).

Chiou u. Wan 2007
Chiou, W.-B. u. Wan, C.-S. (2007): „The dynamic change of self-efficacy in information searching on the internet: Influence of valence of experience and prior self-efficacy". In: *J. Psychol.* 141 (6). PMID: 18044273, S. 589–603 (siehe S. 37, 44, 91, 99, 344).

Churchill u. Surprenant 1982
Churchill JR., G. A. u. Surprenant, C. (1982): „An investigation into the determinants of customer satisfaction". In: *J. Mark. Res.* 19 (4), S. 491–504 (siehe S. 65).

Clemmensen u. Borlund 2016
Clemmensen, M. L. u. Borlund, P. (2016): „Order effect in interactive information retrieval evaluation: An empirical study". In: *J. Doc.* 72 (2), S. 194–213 (siehe S. 122).

Cohen 1960
Cohen, J. (1960): „A coefficient of agreement for nominal scales". In: *Educ. Psychol. Meas.* 20 (1), S. 37–46 (siehe S. 103).

Cohen et al. 1999
Cohen, P.; Cohen, J.; Aiken, L. S. u. West, S. G. (1999): „The problem of units and the circumstance for POMP". In: *Multivariate Behav. Res.* 34 (3), S. 315–346 (siehe S. 199).

Compeau u. Higgins 1995
Compeau, D. R. u. Higgins, C. A. (1995): „Computer self-efficacy: Development of a measure and initial test". In: *Manag. Inf. Syst. Q.* 19 (2), S. 189–211 (siehe S. 36).

Conrath u. Mignen 1990
Conrath, D. W. u. Mignen, O. P. (1990): „What is being done to measure user satisfaction with EDP/MIS“. In: *Inform. Manage.* 19 (1), S. 7–19 (siehe S. 14).

Cox u. Fisher 2004
Cox, A. u. Fisher, M. (2004): „Expectation as a mediator of user satisfaction“. In: *Proc. 13th WWW: Workshop on measuring web search effectiveness: The user perspective* (New York, 17.–22. Mai 2004). New York: ACM, S. 1–6 (siehe S. 22 ff., 44, 71 ff., 86, 93, 108, 325).

Cuadra u. Katter 1967
Cuadra, C. S. u. Katter, R. V. (1967): „Opening the black box of relevance“. In: *J. Doc.* 23 (4), S. 291–303 (siehe S. 4, 21, 108 f., 111 f.).

Deci u. Ryan 1993
Deci, E. L. u. Ryan, R. M. (1993): „Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik“. In: *Z. f. Päd.* 39 (2), S. 223–238 (siehe S. 32).

Doll u. Torkzadeh 1988
Doll, W. J. u. Torkzadeh, G. (1988): „The measurement of end-user computing satisfaction“. In: *Manag. Inf. Syst. Q.* 12 (2), S. 259–274 (siehe S. 118 f., 167 f., 199 f., 207, 279, 282, 401, 414, 460).

Doll u. Xia 1997
Doll, W. J. u. Xia, W. (1997): „Confirmatory factor analysis of the end-user computing satisfaction instrument: A replication“. In: *J. End User Comput.* 9 (2), S. 24–32 (siehe S. 119, 322).

Doll et al. 1994
Doll, W. J.; Xia, W. u. Torkzadeh, G. (1994): „A confirmatory factor analysis of the end-user computing satisfaction instrument“. In: *Manag. Inf. Syst. Q.* 18 (4), S. 453–461 (siehe S. 119, 322).

Dommes et al. 2011
Dommes, A.; Chevalier, A. u. Lia, S. (2011): „The role of cognitive flexibility and vocabulary abilities of younger and older users in searching for information on the web“. In: *Appl. Cognitive Psych.* 25 (5), S. 717–726 (siehe S. 38 f., 45).

Dong et al. 2005
Dong, P.; Loh, M. u. Mondry, A. (2005): „Relevance similarity: An alternative means to monitor information retrieval systems“. In: *Biomed. Digit. Libr.* 2 (1) (siehe S. 30, 125 f.).

Dostert u. Kelly 2009
Dostert, M. u. Kelly, D. (2009): „Users‘ stopping behaviors and estimates of recall“. In: *Proc. 32nd SIGIR* (Boston, 19.–23. Juli 2009). New York: ACM, S. 820–821 (siehe S. 63, 86, 323).

Edwards u. Kelly 2016
Edwards, A. u. Kelly, D. (2016): „How does interest in a work task impact search behavior and engagement?“ In: *Proc. CHIIR* (Carrboro, 13.–17. März 2016). New York: ACM, S. 249–252 (siehe S. 34).

Eisenberg 1988
Eisenberg, M. B. (1988): „Measuring relevance judgments“. In: *Inf. Process. Manage.* 24 (4), S. 373–389 (siehe S. 112).

Eisenberg u. Barry 1988
Eisenberg, M. u. Barry, C. (1988): „Order effects: A study of the possible influence of presentation order on user judgments of document relevance“. In: *J. Am. Soc. Inf. Sci.* 39 (5), S. 293–300 (siehe S. 24, 112, 122 f.).

Eisenberg u. Hu 1987
Eisenberg, M. u. Hu, X. (1987): „Dichotomous relevance judgments and the evaluation of information systems“. In: *Proc. 50th ASIS* (Boston, 4.–8. Okt. 1987). Bd. 24. Medford: Learned Information, S. 66–70 (siehe S. 112 f.).

Ellis 1989
Ellis, D. (1989): „A behavioural approach to information retrieval system design“. In: *J. Doc.* 45 (3), S. 171–212 (siehe S. 8).

Embi 2007
Embi, R. (2007): „Computer anxiety and computer self-efficacy among accounting educators at Universiti Teknologi MARA, Malaysia“. Diss. Blacksburg: Virginia Tech (siehe S. 36).

Enochsson 2005
Enochsson, A. (2005): „A gender perspective on internet use: Consequences for information seeking“. In: *Inform. Res.* 10 (4), S. 1–14 (siehe S. 41).

Festinger 1978
Festinger, L. (1978): *Theorie der kognitiven Dissonanz*. Bern: Huber (siehe S. 67).

Fidel u. Crandall 1997
Fidel, R. u. Crandall, M. (1997): „Users' perception of the performance of a filtering system". In: *Proc. 20th SIGIR* (Philadelphia, 27.–31. Juli 1997). New York: ACM, S. 198–205 (siehe S. 50, 85).

Field et al. 2012
Field, A.; Miles, J. u. Field, Z. (2012): *Discovering statistics using R*. London: SAGE (siehe S. 128–131, 135 f., 138, 201–204, 280).

Fitzgerald u. Galloway 2001
Fitzgerald, M. A. u. Galloway, C. (2001): „Relevance judging, evaluation, and decision making in virtual libraries: A descriptive study". In: *J. Am. Soc. Inf. Sci. Technol.* 52 (12), S. 989–1010 (siehe S. 50, 85).

Flavián-Blanco et al. 2011
Flavián-Blanco, C.; Gurrea-Sarasa, R. u. Orús-Sanclemente, C. (2011): „Analyzing the emotional outcomes of the online search behavior with search engines". In: *Comput. Human Behav.* 27 (1), S. 540–551 (siehe S. 35).

Folkes 1984
Folkes, V. S. (1984): „Consumer reactions to product failure: An attributional approach". In: *J. Consum. Res.* 10 (4), S. 398–409 (siehe S. 68).

Folkes et al. 1987
Folkes, V. S.; Koletsky, S. u. Graham, J. L. (1987): „A field study of causal inferences and consumer reaction: The view from the airport". In: *J. Consum. Res.* 13 (4), S. 534–539 (siehe S. 68).

Fox u. Weisberg 2011
Fox, J. u. Weisberg, S. (2011): *An R companion to applied regression*. 2. Aufl. Thousand Oaks: SAGE. URL: `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion` (siehe S. 139).

Fraillon et al. 2014
Fraillon, J.; Ainley, J.; Schulz, W.; Friedman, T. u. Gebhardt, E. (2014): *Preparing for life in a digital age: The IEA international computer and information literacy study*. Cham: Springer (siehe S. 42).

Freund u. Wildemuth 2014
Freund, L. u. Wildemuth, B. M. (2014): „Documenting and studying the use of assigned search tasks: RepAST". In: *Proc. Assoc. Info. Sci. Tech.* 51 (1), S. 1–4 (siehe S. 98, 100).

Frøkjær et al. 2000
Frøkjær, E.; Hertzum, M. u. Hornbæk, K. (2000): „Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated?" In: *Proc. 18th CHI* (Den Haag, 1.–6. Apr. 2000). New York: ACM, S. 345–352 (siehe S. 47).

Fuhr 2011
Fuhr, N. (2011): *Einführung in Information Retrieval: Skriptum zur Vorlesung im SS 11*. URL: `http://www.is.informatik.uni-duisburg.de/courses/ir_ss11/folien/skript_1-6.pdf` (verifiziert am 20. 05. 2011) (siehe S. 2).

Gediga et al. 2005
Gediga, G.; Gildhorn, A. u. Colver, B. (2005): *Evaluation von vascoda.de aus Benutzersicht: Ergebnisse der Nutzerbefragung*. URL: `http://www.dl-forum.de/dateien/Evaluation_vascoda_Ergebnisse_Befragung_2005.pdf` (verifiziert am 10. 06. 2008) (siehe S. 145).

Gluck 1995
Gluck, M. (1995): „Understanding performance in information systems: Blending relevance and competence". In: *J. Am. Soc. Inf. Sci.* 46. July, S. 446–460 (siehe S. 28).

Granka et al. 2004
Granka, L. A.; Joachims, T. u. Gay, G. K. (2004): „Eye-tracking analysis of user behavior in WWW search". In: *Proc. 27th SIGIR* (Sheffield, 25.–29. Juli 2004). New York: ACM, S. 478–479 (siehe S. 91).

Greisdorf u. Spink 2001
Greisdorf, H. u. Spink, A. (2001): „Median measure: An approach to IR systems evaluation". In: *Inf. Process. Manage.* 37 (6), S. 843–857 (siehe S. 112 f.).

Greve u. Wentura 1997
Greve, W. u. Wentura, D. (1997): *Wissenschaftliche Beobachtung: Eine Einführung*. 2. Aufl. Weinheim: Beltz (siehe S. 103, 238 f.).

Guo et al. 2011
Guo, Q.; White, R. W.; Zhang, Y.; Anderson, B. u. Dumais, S. T. (2011): „Why searchers switch: Understanding and predicting engine switching rationales“. In: *Proc. 34th SIGIR* (Beijing, 24.–28. Juli 2011). New York: ACM, S. 335–344 (siehe S. 5, 25).

Habel et al. 2016
Habel, J.; Alavi, S.; Schmitz, C.; Schneider, J.-V. u. Wieseke, J. (2016): „When Do Customers Get What They Expect? Understanding the Ambivalent Effects of Customers' Service Expectations on Satisfaction“. In: *J. Serv. Res.* 19 (4), S. 361–379 (siehe S. 210, 318).

Hargittai u. Shafer 2006
Hargittai, E. u. Shafer, S. (2006): „Differences in actual and perceived online skills: The role of gender“. In: *Soc. Sci. Q.* 87 (2), S. 432–448 (siehe S. 41).

Harter 1992
Harter, S. P. (1992): „Psychological relevance and information science“. In: *J. Am. Soc. Inf. Sci.* 43 (9), S. 602–615 (siehe S. 48 ff., 53).

Heath u. White 2008
Heath, A. P. u. White, R. W. (2008): „Defection detection: Predicting search engine switching“. In: *Proc. 17th WWW* (Beijing, 21.–25. Apr. 2008). New York: ACM, S. 1173–1174 (siehe S. 25 f.).

Heckhausen u. Heckhausen 2010
Heckhausen, J. u. Heckhausen, H. (2010): *Motivation und Handeln*. Berlin: Springer (siehe S. 173).

Heider 1958
Heider, F. (1958): *The psychology of interpersonal relations*. New York: Wiley (siehe S. 68).

Heinström 2006
Heinström, J. (2006): „Fast surfing for availability or deep diving into quality - motivation and information seeking among middle and high school students“. In: *Inform. Res.* 11 (4) (siehe S. 32 ff.).

Hersh et al. 2000
Hersh, W.; Turpin, A.; Price, S.; Chan, B.; Kramer, D.; Sacherek, L. u. Olson, D. (2000): „Do batch and user evaluations give the same results?“ In: *Proc. 23rd SIGIR* (Athen, 24.–28. Juli 2000). New York: ACM, S. 17–24 (siehe S. 43, 57, 101, 111, 335).

Hill et al. 2011
Hill, R. L.; Dickinson, A.; Arnott, J. L.; Gregor, P. u. McIver, L. (2011): „Older web users' eye movements: experience counts“. In: *Proc. 29th CHI* (Vancouver, 7.–12. Mai 2011). New York: ACM, S. 1151–1160 (siehe S. 39 f., 91).

Hlavac 2015
Hlavac, M. (2015): *stargazer: Well-formatted regression and summary statistics tables*. R package version 5.2. Cambridge. URL: `http://CRAN.R-project.org/package=stargazer` (verifiziert am 25. 06. 2017) (siehe S. 139).

Hohlfeld et al. 2013
Hohlfeld, T. N.; Ritzhaupt, A. D. u. Barron, A. E. (2013): „Are gender differences in perceived and demonstrated technology literacy significant? It depends on the model“. In: *Educ. Technol. Res. Dev.* 61 (4), S. 639–663 (siehe S. 42).

Hölscher 2000
Hölscher, C. (2000): „Informationssuche im Internet: Web-Expertise und Wissenseinflüsse“. Diss. Freiburg: Albert-Ludwigs-Universität (siehe S. 126, 172).

Hölscher u. Strube 2000
Hölscher, C. u. Strube, G. (2000): „Web search behavior of Internet experts and newbies“. In: *Comput. Netw.* 33 (1–6), S. 337–346 (siehe S. 28 f., 91, 95 f., 125 f.).

Homburg 2012
Homburg, C., Homburg, C. (Hrsg.): (2012): *Kundenzufriedenheit: Konzepte - Methoden - Erfahrungen*. 8. Aufl. Wiesbaden: Gabler (siehe S. 65).

Homburg u. Rudolph 1997
Homburg, C. u. Rudolph, B. (1997): „Theoretische Perspektiven zur Kundenzufriedenheit“. In: Simon, H. u. Homburg, C., Simon, H. u. Homburg, C. (Hrsg.): *Kundenzufriedenheit: Konzepte — Methoden — Erfahrungen*. Wiesbaden: Gabler, S. 31–51 (siehe S. 66).

Homburg u. Stock-Homburg 2012
Homburg, C. u. Stock-Homburg, R. (2012): „Theoretische Perspektiven zur Kundenzufriedenheit“. In:

Homburg, C., Homburg, C. (Hrsg.): *Kundenzufriedenheit: Konzepte - Methoden - Erfahrungen*. 8. Aufl. Wiesbaden: Gabler, S. 17–52 (siehe S. 65–68).

Horn u. Cattell 1967
Horn, J. L. u. Cattell, R. B. (1967): „Age differences in fluid and crystallized intelligence". In: *Acta Psychol.* 26, S. 107–129 (siehe S. 38).

Hovland et al. 1957
Hovland, C. I.; Harvey, O. J. u. Sherif, M. (1957): „Assimilation and contrast effects in reactions to communication and attitude change". In: *J. Abnorm. Soc. Psychol.* 55 (2), S. 244–252 (siehe S. 67).

Howard 1994
Howard, D. L. (1994): „Pertinence as reflected in personal constructs". In: *J. Am. Soc. Inf. Sci.* 45 (3), S. 172–185 (siehe S. 51, 85).

Hu et al. 2011
Hu, V.; Stone, M.; Pedersen, J. u. White, R. W. (2011): „Effects of search success on search engine re-use". In: *Proc. 20th CIKM* (Glasgow, 24.–28. Okt. 2011). New York: ACM, S. 1841–1846 (siehe S. 26, 82, 118, 120).

Huang u. Wang 2004
Huang, M.-h. u. Wang, H.-y. (2004): „The influence of document presentation order and number of documents judged on users' judgments of relevance". In: *J. Am. Soc. Inf. Sci. Technol.* 55 (11), S. 970–979 (siehe S. 24, 111 f., 122 f.).

Huffman u. Hochster 2007
Huffman, S. B. u. Hochster, M. (2007): „How well does result relevance predict session satisfaction". In: *Proc. 30th SIGIR* (Amsterdam, 23.–27. Juli 2007). New York: ACM, S. 567–574 (siehe S. 77, 86, 97, 118, 147).

Igbaria u. Iivari 1995
Igbaria, M. u. Iivari, J. (1995): „The effects of self-efficacy on computer usage". In: *Omega* 23 (6), S. 587–605 (siehe S. 36).

Ingwersen u. Järvelin 2005
Ingwersen, P. u. Järvelin, K. (2005): *The turn: Integration of information seeking and retrieval in context*. Dordrecht: Springer (siehe S. 2 f.).

Irle 2017
Irle, G. J. (2017): „Gefühlserleben bei der Informationssuche im Internet: Eine qualitative Studie zur Individualität und Alltäglichkeit der Sucherfahrung". Diss. Hildesheim: Universität Hildesheim (siehe S. 34).

Ives et al. 1983
Ives, B.; Olson, M. H. u. Baroudi, J. J. (1983): „The measurement of user information satisfaction". In: *Commun. ACM* 26 (10), S. 785–793 (siehe S. 118 f.).

Janes 1991a
Janes, J. W. (1991a): „Relevance judgments and the incremental presentation of document representations". In: *Inf. Process. Manage.* 27 (6), S. 629–646 (siehe S. 112 f.).

Janes 1991b
Janes, J. W. (1991b): „The binary nature of continuous relevance judgments: A study of users' perceptions". In: *J. Am. Soc. Inf. Sci. Technol.* 42 (10), S. 754–756 (siehe S. 112 f.).

Jansen u. Spink 2006
Jansen, B. J. u. Spink, A. (2006): „How are we searching the World Wide Web? A comparison of nine search engine transaction logs". In: *Inf. Process. Manage.* 42 (1), S. 248–263 (siehe S. 90).

Jansen et al. 2000
Jansen, B. J.; Spink, A. u. Saracevic, T. (2000): „Real life, real users, and real needs: a study and analysis of user queries on the web". In: *Inf. Process. Manage.* 36 (2), S. 207–227 (siehe S. 90, 105).

Jansen et al. 2007
Jansen, B. J.; Zhang, M. u. Zhang, Y. (2007): „The effect of brand awareness on the evaluation of search engine results". In: *Proc. 25th CHI* (San Jose, 30. Apr.–3. Mai 2007). New York: ACM, S. 2471–2476 (siehe S. 1, 4, 24 f., 71, 86, 108 f., 118).

Järvelin 2009
Järvelin, K. (2009): „Explaining user performance in information retrieval: Challenges to IR evaluation". In: *Proc. 2nd ICTIR* (Cambridge, 10.–12. Sep. 2009). Berlin: Springer, S. 289–296 (siehe S. 116 f.).

Järvelin u. Ingwersen 2004
Järvelin, K. u. Ingwersen, P. (2004): „Information seeking research needs extension towards tasks and technology“. In: *Inform. Res.* 10 (1) (siehe S. 56).

Järvelin u. Kekäläinen 2000
Järvelin, K. u. Kekäläinen, J. (2000): „IR evaluation methods for retrieving highly relevant documents“. In: *Proc. 23rd SIGIR* (Athen, 24.–28. Juli 2000). New York: ACM, S. 41–48 (siehe S. 3, 106).

Järvelin u. Kekäläinen 2002
Järvelin, K. u. Kekäläinen, J. (2002): „Cumulated gain-based evaluation of IR techniques“. In: *ACM Trans. Inf. Syst.* 20 (4), S. 422–446 (siehe S. 3, 106).

Jenkins et al. 2003
Jenkins, C.; Corritore, C. L. u. Wiedenbeck, S. (2003): „Patterns of information seeking on the web: A qualitative study of domain expertise and web expertise“. In: *IT & Soc.* 1 (3), S. 64–89 (siehe S. 28 f., 95, 125 f.).

Jiang u. Allan 2016
Jiang, J. u. Allan, J. (2016): „Correlation between system and user metrics in a session“. In: *Proc. CHIIR* (Carrboro, 13.–17. März 2016). New York: ACM, S. 285–288 (siehe S. 64, 80, 86).

Jiang et al. 2015
Jiang, J.; Hassan Awadallah, A.; Shi, X. u. White, R. W. (2015): „Understanding and predicting graded search satisfaction“. In: *Proc. 8th WSDM* (Shanghai, 2.–6. Feb. 2015). New York: ACM, S. 57–66 (siehe S. 82 f., 87).

Jiang et al. 2017
Jiang, J.; He, D.; Kelly, D. u. Allan, J. (2017): „Understanding ephemeral state of relevance“. In: *Proc. CHIIR* (Oslo, 7.–11. März 2017). New York: ACM, S. 137–146 (siehe S. 82 f., 87, 345).

Joachims et al. 2005
Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H. u. Gay, G. (2005): „Accurately interpreting clickthrough data as implicit feedback“. In: *Proc. 28th SIGIR* (Salvador, 15.–19. Aug. 2005). New York: ACM, S. 154–161 (siehe S. 5, 24, 347).

Joachims et al. 2007
Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; Radlinski, F. u. Gay, G. (2007): „Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search“. In: *ACM Trans. Inf. Syst.* 25 (2), S. 1–26 (siehe S. 24 f.).

Johnson et al. 2003
Johnson, F. C.; Griffiths, J. R. u. Hartley, R. J. (2003): „Task dimensions of user evaluations of information retrieval systems“. In: *Inform. Res.* 8 (4) (siehe S. 75, 86, 323).

Johnson et al. 1995
Johnson, M. D.; Anderson, E. W. u. Fornell, C. (1995): „Rational and adaptive performance expectations in a customer satisfaction framework“. In: *J. Consum. Res.* 21 (4), S. 695–707 (siehe S. 66).

Jonkisz et al. 2008
Jonkisz, E.; Moosbrugger, H. u. Brandt, H. (2008): „Planung und Entwicklung von psychologischen Tests und Fragebogen“. In: Moosbrugger, H. u. Kelava, A., Moosbrugger, H. u. Kelava, A. (Hrsg.): *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer, S. 27–72 (siehe S. 55, 124 f.).

Juan u. Chang 2005
Juan, Y.-F. u. Chang, C.-C. (2005): „An analysis of search engine switching behavior using click streams“. In: *Proc. 14th WWW* (Chiba, 10.–14. Mai 2005). New York: ACM, S. 1050–1051 (siehe S. 26).

Kaczmirek 2003
Kaczmirek, L. (2003): „Information und Selektion: Gebrauchstauglichkeit der Ergebnisseiten von Suchmaschinen“. Dipl.-Arb. Mannheim: Universität Mannheim (siehe S. 145, 147 f.).

Kaiser 2005
Kaiser, M.-O. (2005): *Erfolgsfaktor Kundenzufriedenheit: Dimensionen und Messmöglichkeiten*. Bd. 64. Berlin: ESV (siehe S. 65, 67, 69 f.).

Käki u. Aula 2008
Käki, M. u. Aula, A. (2008): „Controlling the complexity in comparing search user interfaces via user studies“. In: *Inf. Process. Manage.* 44 (1), S. 82–91 (siehe S. 93 f., 97, 99, 114, 116).

Karanam u. van Oostendorp 2016
Karanam, S. u. van Oostendorp, H. (2016): „Age-related differences in the content of search queries

when reformulating“. In: *Proc. 34th CHI* (San Jose, 7.–12. Mai 2016). New York: ACM, S. 5720–5730 (siehe S. 38 f., 45).

Keane et al. 2008
Keane, M. T.; O'Brien, M. u. Smyth, B. (2008): „Are people biased in their use of search engines?“ In: *Commun. ACM* 51 (2), S. 49–52 (siehe S. 5, 24 f., 347).

Kelly 2009
Kelly, D. (2009): „Methods for evaluating interactive information retrieval systems with users“. In: *Found. Trends Inf. Retr.* 3 (1-2), S. 1–224 (siehe S. 4, 6 f., 91, 99, 102 f., 105, 107, 117 f., 122).

Kelly u. Belkin 2004
Kelly, D. u. Belkin, N. J. (2004): „Display time as implicit feedback: Understanding task effects“. In: *Proc. 27th SIGIR* (Sheffield, 25.–29. Juli 2004). New York: ACM, S. 377–384 (siehe S. 170).

Kelly u. Cool 2002
Kelly, D. u. Cool, C. (2002): „The effects of topic familiarity on information search behavior“. In: *Proc. 2nd JCDL* (Portland, 13. Juli 2002–17. Juli 2001). New York: ACM, S. 74–75 (siehe S. 30 f., 125).

Kelly et al. 2007
Kelly, D.; Fu, X. u. Shah, C. (2007): *Effects of rank and precision of search results on users' evaluations of system performance*. UNC SILS Technical Report TR-2007-02. University of Noth Carolina. URL: `http://sils.unc.edu/sites/default/files/general/research/TR-2007-02.pdf` (verifiziert am 15. 07. 2011) (siehe S. 62 f., 75 f., 78, 86 f., 98 f., 101, 111, 118, 166, 323, 335).

Kelly et al. 2008a
Kelly, D.; Harper, D. J. u. Landau, B. (2008a): „Questionnaire mode effects in interactive information retrieval experiments“. In: *Inf. Process. Manage.* 44 (1), S. 122–141 (siehe S. 73, 86, 124 f.).

Kelly et al. 2008b
Kelly, D.; Shah, C.; Sugimoto, C. R.; Bailey, E. W.; Clemens, R. A.; Irvine, A. K.; Johnson, N. A.; Ke, W.; Oh, S.; Poljakova, A.; Rodriguez, M. A.; van Noord, M. G. u. Zhang, Y. (2008b): *Method bias?: The effects of performance feedback on users' evaluations of an interactive IR system*. UNC SILS Technical Report TR-2008-01. University of Noth Carolina. URL: `http://sils.unc.edu/sites/default/files/general/research/TR_2008_01.pdf` (verifiziert am 15. 07. 2011) (siehe S. 35 f., 73, 86, 124, 169).

Kelly 1955
Kelly, G. A. (1955): *The psychology of personal constructs: Volume 1: A theory of personality*. New York: Norton (siehe S. 51).

Kinzie et al. 1994
Kinzie, M. B.; Delcourt, M. A. B. u. Powers, S. M. (1994): „Computer technologies: Attitudes and self-efficacy across undergraduate disciplines“. In: *Res. High. Educ.* 35 (6), S. 745–768 (siehe S. 36).

Kiseleva et al. 2016
Kiseleva, J.; Williams, K.; Jiang, J.; Hassan Awadallah, A.; Crook, A. C.; Zitouni, I. u. Anastasakos, T. (2016): „Understanding user satisfaction with intelligent assistants“. In: *Proc. CHIIR* (Carrboro, 13.–17. März 2016). New York: ACM, S. 121–130 (siehe S. 35, 79, 87, 335).

Kissel 1995
Kissel, G. V. (1995): „The effect of computer experience on subjective and objective software usability measures“. In: *Proc. 13th CHI* (Denver, 7.–11. Mai 1995). New York: ACM, S. 284–285 (siehe S. 23 f., 125).

Kloke u. McKean 2014
Kloke, J. u. McKean, J. (2014): *npsm: Package for nonparametric statistical methods using R*. R package version 0.5. URL: `https://CRAN.R-project.org/package=npsm` (verifiziert am 25. 06. 2017) (siehe S. 139).

Krapp u. Ryan 2002
Krapp, A. u. Ryan, R. M. (2002): „Selbstwirksamkeit und Lernmotivation: Eine kritische Betrachtung der Theorie von Bandura aus der Sicht der Selbstbestimmungstheorie und der pädagogisch-psychologischen Interessentheorie“. In: *Z. f. Päd.* (44. Beiheft), S. 54–82 (siehe S. 32).

Küchenhoff 2006
Küchenhoff, H. (2006): *Statistik für Kommunikationswissenschaftler*. 2. Aufl. Konstanz: UVK (siehe S. 112).

Kuhlthau 1993a
Kuhlthau, C. C. (1993a): „A principle of uncertainty for information seeking“. In: *J. Doc.* 49 (4), S. 339–355 (siehe S. 8, 54).

Kuhlthau 1993b
Kuhlthau, C. C. (1993b): *Seeking meaning: A process approach to library and information services*. Norwood: Ablex (siehe S. 55).

Kwahk u. Oh 2009
Kwahk, K.-Y. u. Oh, S.-W. (2009): „Examining the effect of user expectations on system use activity". In: *Proc. 17th ECIS* (Verona, 8.–10. Juni 2009). Atlanta: AIS, S. 352–363 (siehe S. 72, 86, 147).

Lamm 2008
Lamm, K. (2008): „Das Confirmation/Diskonfirmation-Paradigma der Kundenzufriedenheit im Kontext des Information Retrieval". Mag.-Arb. Hildesheim: Universität Hildesheim (siehe S. 141 f., 145, 148, 152, 156, 158, 160 f.).

Lamm et al. 2010a
Lamm, K.; Mandl, T.; Womser-Hacker, C. u. Greve, W. (2010a): „The influence of expectation and system performance on user satisfaction with retrieval systems". In: *Proc. 3rd EVIA* (Tokyo, 15. Juni 2010). Tokyo: National Institute of Informatics, S. 60–68 (siehe S. 142).

Lamm et al. 2010b
Lamm, K.; Mandl, T.; Womser-Hacker, C. u. Greve, W. (2010b): „User experiments with search services: Methodological challenges for measuring the perceived quality". In: *Proc. 3rd ISCA/DEGA: Tutorial and Workshop on Perceptual Quality of Systems* (Bautzen, 6.–8. Sep. 2010). Dresden: ISCA/DEGA, S. 64–69 (siehe S. 142, 163).

Lancaster u. Warner 1993
Lancaster, F. W. u. Warner, A. J. (1993): *Information retrieval today*. Arlington: Info. Resources Pr. (siehe S. 117).

Langfeld 2006
Langfeld, H.-P. (2006): *Psychologie für die Schule*. Weinheim: Beltz (siehe S. 32).

Large et al. 2002
Large, A.; Beheshti, J. u. Rahman, T. (2002): „Gender differences in collaborative web searching behavior: An elementary school study". In: *Inf. Process. Manage.* 38 (3), S. 427–443 (siehe S. 40 f.).

Lazonder et al. 2000
Lazonder, A. W.; Biemans, H. J. A. u. Wopereis, I. G. J. H. (2000): „Differences between novice and experienced users in searching information on the World Wide Web". In: *J. Am. Soc. Inf. Sci.* 51 (6), S. 576–581 (siehe S. 28).

Leader u. Klein 1996
Leader, L. F. u. Klein, J. D. (1996): „The effects of search tool type and cognitive style on performance during hypermedia database searches". In: *Educ. Technol. Res. Dev.* 44 (2), S. 5–15 (siehe S. 20).

Lewandowski 2014
Lewandowski, D. (2014): „Wie lässt sich die Zufriedenheit der Suchmaschinennutzer mit ihren Suchergebnissen erklären". In: Krah, H. u. Müller-Terpitz, R., Krah, H. u. Müller-Terpitz, R. (Hrsg.): *Suchmaschinen*. Bd. 4. Passauer Schriften zur interdisziplinären Medienforschung 4. Münster: LIT, S. 35–52 (siehe S. 73).

Li et al. 2011
Li, Y.; Chen, Y.; Liu, J.; Cheng, Y.; Wang, X.; Chen, P. u. Wang, Q. (2011): „Measuring task complexity in information search from user's perspective". In: *Proc. 74th ASIST* (New Orleans, 9.–12. Okt. 2011). Bd. 48. Silver Spring: ASIST, S. 1–8 (siehe S. 43 f.).

Lindblom et al. 2012
Lindblom, K.; Gregory, T.; Wilson, C.; Flight, I. H. K. u. Zajac, I. (2012): „The impact of computer self-efficacy, computer anxiety, and perceived usability and acceptability on the efficacy of a decision support tool for colorectal cancer screening". In: *J. Am. Med. Inform. Assn.* 19 (3), S. 407–412 (siehe S. 36).

Liu u. Wei 2016
Liu, C. u. Wei, Y. (2016): „The impacts of time constraint on users' search strategy during search process". In: *Proc. 79th ASIST* (Kopenhagen, 14.–18. Okt. 2016). Silver Springs: ASIST, S. 1–9 (siehe S. 44).

Luhmann 2013
Luhmann, M. (2013): *R für Einsteiger: Einführung in die Statistiksoftware für die Sozialwissenschaften*. 3. Aufl. Weinheim: Beltz (siehe S. 128, 204, 279).

Luo et al. 2017
Luo, C.; Li, X.; Liu, Y.; Sakai, T.; Zhang, F.; Zhang, M. u. Ma, S. (2017): „Investigating users' time perception during web search". In: *Proc. 35th CHI* (Denver, 6.–11. Mai 2017). New York: ACM, S. 127–136 (siehe S. 35, 79, 87, 93, 335, 344).

Maglaughlin u. Sonnenwald 2002
Maglaughlin, K. L. u. Sonnenwald, D. H. (2002): „User perspectives on relevance criteria: A comparison among relevant, partially relevant , and not-relevant judgements". In: *J. Am. Soc. Inf. Sci. Technol.* 53 (5), S. 327–342 (siehe S. 50, 85).

Mandl 2006
Mandl, T. (2006): „Die automatische Bewertung der Qualität von Internet- Seiten im Information Retrieval". Habil.-Schr. Hildesheim: Universität Hildesheim (siehe S. 2).

Mandl 2008
Mandl, T. (2008): „Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance". In: *Informatica* 32 (1), S. 27–38 (siehe S. 2).

McHaney et al. 1999
McHaney, R.; Hightower, R. u. White, D. (1999): „EUCS test-retest reliability in representational model decision support systems". In: *Inform. Manage.* 36 (2), S. 109–119 (siehe S. 119).

Moghadasi et al. 2013
Moghadasi, S. I.; Ravana, S. D. u. Raman, S. N. (2013): „Low-cost evaluation techniques for information retrieval systems: A review". In: *J. Informetr.* 7 (2), S. 301–312 (siehe S. 106).

Monoi et al. 2005
Monoi, S.; O'Hanlon, N. u. Diaz, K. R. (2005): „Online searching skills: Development of an inventory to assess self-efficacy". In: *J. Acad. Librariansh.* 31 (2), S. 98–105 (siehe S. 36 f., 344).

Moosbrugger u. Kelava 2011
Moosbrugger, H. u. Kelava, A. (2011): *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer (siehe S. 102, 112 f.).

Mulder u. Yaar 2007
Mulder, S. u. Yaar, Z. (2007): *The user is always right: A practical guide to creating and using personas for the web*. Berkeley: New Riders (siehe S. 13).

Neugebauer u. Porst 2001
Neugebauer, B. u. Porst, R. (2001): *Patientenzufriedenheit: Ein Literaturbericht*. ZUMA-Methodenbericht 7. Zentrum für Umfragen, Methoden und Analysen. URL: `http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2001/01_07.pdf` (verifiziert am 09. 07. 2015) (siehe S. 124).

O'Brien u. Keane 2006
O'Brien, M. u. Keane, M. T. (2006): „Modeling result-list searching in the World Wide Web: The role of relevance topologies and trust bias". In: *Proc. 28th CogSci* (Vancouver, 26.–29. Juli 2006). Hillsdale: Erlbaum, S. 1881–1886 (siehe S. 24 f.).

Olson u. Dover 1979
Olson, J. C. u. Dover, P. A. (1979): „Disconfirmation of consumer expectations through product trial". In: *J. Appl. Psychol.* 64 (2), S. 179–189 (siehe S. 83, 87).

Omar u. Lascu 1993
Omar, M. H. u. Lascu, D.-N. (1993): „Development of a user information satisfaction scale: An alternative measure with wide applicability". In: *J. Inform. Technol. Manag.* 4 (2), S. 1–13 (siehe S. 118 ff.).

Osgood 1962
Osgood, C. E. (1962): „Studies on the generality of affective meaning systems". In: *Am. Psychol.* 17 (1), S. 10–28 (siehe S. 119).

Pajares 1997
Pajares, F. (1997): „Current directions in self-efficacy research". In: Maehr, M. u. Pintrich, P. R., Maehr, M. u. Pintrich, P. R. (Hrsg.): *Advances in motivation and achievment*. 10. Aufl. Greenwich: JAI Press, S. 1–49 (siehe S. 35).

Pak u. Price 2008
Pak, R. u. Price, M. M. (2008): „Designing an information search interface for younger and older adults". In: *Hum. Factors* 50 (4), S. 614–628 (siehe S. 38).

Palmquist u. Kim 1998
Palmquist, R. A. u. Kim, K.-S. (1998): „Modeling the users of information systems: Some theories and methods“. In: *Ref. Libr.* 18 (60), S. 3–25 (siehe S. 4, 91).

Palmquist u. Kim 2000
Palmquist, R. A. u. Kim, K.-S. (2000): „Cognitive style and on-line database search experience as predictors of Web search performance“. In: *J. Am. Soc. Inf. Sci.* 51 (6), S. 558–566 (siehe S. 20 f., 27 f., 125).

Pan et al. 2004
Pan, B.; Hembrooke, H. A.; Gay, G. K.; Granka, L. A.; Feusner, M. K. u. Newman, J. K. (2004): „The determinants of web page viewing behavior: An eye-tracking study“. In: *Proc. ETRA* (San Antonio, 22.–24. März 2004). New York: ACM, S. 147–154 (siehe S. 91).

Pan et al. 2007
Pan, B.; Hembrooke, H. A.; Joachims, T.; Lorigo, L.; Gay, G. K. u. Granka, L. A. (2007): „In Google we trust: Users' decisions on rank, position, and relevance“. In: *J. Comput. Mediat. Commun.* 12 (3), S. 801–823 (siehe S. 5, 24 f., 91, 95, 347).

Parasuraman et al. 1988
Parasuraman, A.; Zeithaml, V. u. Berry, L. (1988): „SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality“. In: *J. Retailing* 64 (1), S. 12–40 (siehe S. 118).

Park et al. 2005
Park, S.; Lee, J. H. u. Bae, H. J. (2005): „End user searching: A Web log analysis of NAVER, a korean web search engine“. In: *Libr. Inf. Sci. Res.* 27 (2), S. 203–221 (siehe S. 90).

Park 1993
Park, T. K. (1993): „The nature of relevance in information retrieval: An empirical study“. In: *Libr. Q.* 63 (3), S. 318–351 (siehe S. 2, 4, 50, 85).

Pham et al. 2010
Pham, M. T.; Goukens, C.; Lehmann, D. R. u. Stuart, J. A. (2010): „Shaping customer satisfaction through self-awareness cues“. In: *J. Mark. Res.* 47 (5), S. 920–932 (siehe S. 68).

Punter et al. 2016
Punter, R. A.; Meelissen, M. R. M. u. Glas, C. A. W. (2016): „Gender differences in computer and information literacy: An exploration of the performances of girls and boys in ICILS 2013“. In: *Eur. Educ. Res. J.* S. 1–19 (siehe S. 42).

Purgailis Parker u. Johnson 1990
Purgailis Parker, L. M. u. Johnson, R. E. (1990): „Does order of presentation affect users' judgment of documents?“ In: *J. Am. Soc. Inf. Sci. Technol.* 41 (7), S. 493–494 (siehe S. 102, 111, 122 f.).

R Core Team 2017
R Core Team (2017): *R: A language and environment for statistical computing*. URL: `https://www.R-project.org/` (verifiziert am 25. 06. 2017) (siehe S. 138 f.).

Raab-Steiner u. Benesch 2010
Raab-Steiner, E. u. Benesch, M. (2010): *Der Fragebogen: von der Forschungsidee zur SPSS/PASW-Auswertung*. 2. Aufl. UTB. Wien: Facultas (siehe S. 55).

Radlinski et al. 2008
Radlinski, F.; Kurup, M. u. Joachims, T. (2008): „How does clickthrough data reflect retrieval quality?“ In: *Proc. 17th CIKM* (Napa Valley, 26.–30. Okt. 2008). New York: ACM, S. 43–52 (siehe S. 116).

Resnick u. Lergier 2003
Resnick, M. L. u. Lergier, R. (2003): „Task specific user strategies in on-line search“. In: *J. E-Bus.* 3 (1), S. 1–22 (siehe S. 146).

Revelle 2016
Revelle, W. (2016): *psych: Procedures for psychological, psychometric, and personality research*. R package version 1.6.12. URL: `https://CRAN.R-project.org/package=psych` (verifiziert am 25. 06. 2017) (siehe S. 139).

Richman et al. 1999
Richman, W. L.; Kiesler, S.; Weisband, S. u. Drasgow, F. (1999): „A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews“. In: *J. Appl. Psychol.* 84 (5), S. 754–775 (siehe S. 125).

Richter 2005
Richter, M. (2005): *Dynamik von Kundenerwartungen im Dienstleistungsprozess: Konzeptionalisierung und empirische Befunde*. 1. Aufl. Bd. 17. Basler Schriften zum Marketing. Wiesbaden: Gabler (siehe S. 15 ff., 19, 65, 69).

Richter et al. 2001a
Richter, T.; Naumann, J. u. Horz, H. (2001a): „Computer Literacy, computerbezogene Einstellungen und Computernutzung bei männlichen und weiblichen Studierenden". In: *Proc. Mensch & Computer* (Bad Honnef, 5.–8. März 2001). Wiesbaden: Vieweg+Teubner, S. 71–80 (siehe S. 40 ff.).

Richter et al. 2001b
Richter, T.; Neumann, J. u. Groeben, N. (2001b): „Das Inventar zur Computerbildung (INCOBI): Ein Instrument zur Erfassung von Computer Literacy und computerbezogenen Einstellungen bei Studierenden der Geistes- und Sozialwissenschaften". In: *Psychol. Erz. Unterr.* 48 (1), S. 1–13 (siehe S. 125 f., 172).

Riding u. Cheema 1991
Riding, R. u. Cheema, I. (1991): „Cognitive styles—an overview and integration". In: *Educ. Psychol.* 11 (3-4), S. 193–215 (siehe S. 20).

Rieh u. Xie 2001
Rieh, S. Y. u. Xie, H. (2001): „Patterns and sequences of multiple query reformulations in web searching: A preliminary study". In: *Proc. 64th ASIST* (Washington, 4.–8. Nov. 2001). Medford: Information Today, S. 246–255 (siehe S. 90).

Roy u. Chi 2003
Roy, M. u. Chi, M. T. H. (2003): „Gender differences in patterns of searching the web". In: *J. Educ. Comp. Res.* 29 (3), S. 335–348 (siehe S. 40 f., 91).

Roy et al. 2003
Roy, M.; Taylor, R. u. Chi, M. T. H. (2003): „Searching for information on-line and off-line: Gender differences among middle school students". In: *J. Educ. Comp. Res.* 29 (2), S. 229–252 (siehe S. 40 f.).

Rummel 2014
Rummel, B. (2014): „Probability plotting: A tool for analyzing task completion times". In: *JUS* 9 (4), S. 152–172 (siehe S. 170).

Rushinek u. Rushinek 1986
Rushinek, A. u. Rushinek, S. F. (1986): „What makes users happy?" In: *Commun. ACM* 29 (7), S. 594–598 (siehe S. 14).

Ryan u. Deci 2000
Ryan, R. M. u. Deci, E. L. (2000): „Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being". In: *Am. Psychol.* 55 (1), S. 68–78 (siehe S. 32 f., 342).

Ryker et al. 1997
Ryker, R.; Nath, R. u. Henson, J. (1997): „Determinants of computer user expectations and their relationships with user satisfaction: an empirical study". In: *Inf. Process. Manage.* 33 (4), S. 529–537 (siehe S. 25).

Sakai 2007
Sakai, T. (2007): „Alternatives to bpref". In: *Proc. 30th SIGIR* (Amsterdam, 23.–27. Juli 2007). New York: ACM, S. 71–78 (siehe S. 106).

Sakai u. Kando 2008
Sakai, T. u. Kando, N. (2008): „On information retrieval metrics designed for evaluation with incomplete relevance assessments". In: *Inf. Retr.* 11 (5), S. 447–470 (siehe S. 106).

Sam et al. 2005
Sam, H. K.; Othm, A. E. A. u. Nordin, Z. S. (2005): „Computer Self-Efficacy, Computer Anxiety, and Attitudes toward the Internet: A Study among Undergraduates in Unimas". In: *J. Educ. Techno. Soc.* 8 (4), S. 205–219 (siehe S. 36).

Sanderson u. Zobel 2005
Sanderson, M. u. Zobel, J. (2005): „Information retrieval system evaluation: effort, sensitivity, and reliability". In: *Proc. 28th SIGIR* (Salvador, 15.–19. Aug. 2005). New York: ACM, S. 162–169 (siehe S. 58).

Sandore 1990
Sandore, B. (1990): „Online Searching: What Measure Satisfaction?" In: *Libr. Inf. Sci. Res.* 12 (1), S. 33–54 (siehe S. 79 f.).

Saracevic 1975
Saracevic, T. (1975): „Relevance: A review of and a framework for the thinking on the notion in information science“. In: *J. Am. Soc. Inf. Sci.* 26 (6), S. 321–343 (siehe S. 48).

Saracevic 1996
Saracevic, T. (1996): „Relevance reconsidered“. In: *Proc. 2nd CoLIS* (Kopenhagen, 13.–16. Okt. 1996). Kopenhagen: Royal School of Librarianship, S. 201–218 (siehe S. 48 f., 51, 82, 326).

Saracevic 2007a
Saracevic, T. (2007a): „Relevance: A review of the literature and a framework for thinking on the notion in information science“. Part II: Nature and manifestations of relevance. In: *J. Am. Soc. Inf. Sci. Technol.* 58 (13), S. 1915–1933 (siehe S. 48).

Saracevic 2007b
Saracevic, T. (2007b): „Relevance: A review of the literature and a framework for thinking on the notion in information science“. Part III: Behavior and effects of relevance. In: *J. Am. Soc. Inf. Sci. Technol.* 58 (13), S. 2126–2144 (siehe S. 48).

Saracevic u. Kantor 1988a
Saracevic, T. u. Kantor, P. (1988a): „A study of information seeking and retrieving: II: Users, questions, and effectiveness“. In: *J. Am. Soc. Inf. Sci.* 39 (3), S. 177–196 (siehe S. 51, 85).

Saracevic u. Kantor 1988b
Saracevic, T. u. Kantor, P. (1988b): „A study of information seeking and retrieving: III. Searchers, searches, and overlap“. In: *J. Am. Soc. Inf. Sci.* 39 (3), S. 197–216 (siehe S. 4, 51, 85).

Saracevic et al. 1988
Saracevic, T.; Kantor, P.; Chamis, A. Y. u. Trivison, D. (1988): „A study of information seeking and retrieving: I. Background and methodology“. In: *J. Am. Soc. Inf. Sci.* 39 (3), S. 161–176 (siehe S. 51, 85, 111).

Sauerwein 2000
Sauerwein, E. (2000): *Das Kano-Modell der Kundenzufriedenheit: Reliabilität und Validität einer Methode zur Klassifizierung von Produkteigenschaften*. Wiesbaden: Dt. Univ.-Verl. (siehe S. 18).

Schamber 1991
Schamber, L. (1991): „Users' criteria for evaluation in a multimedia environment“. In: *Proc. 54th ASIS* (Washington, 27.–31. Okt. 1991). Medford: Learned Information, S. 126–133 (siehe S. 4, 50, 85).

Schamber 1994
Schamber, L. (1994): „Relevance and information behavior“. In: *Annu. Rev. Inform. Sci. Technol.* 29, S. 3–48 (siehe S. 48, 50, 85).

Schamber et al. 1990
Schamber, L.; Eisenberg, M. u. Nilan, M. S. (1990): „A re-examination of relevance: Toward a dynamic, situational definition“. In: *Inf. Process. Manage.* 26 (6), S. 755–776 (siehe S. 49 f., 329).

Scharnbacher u. Kiefer 1996
Scharnbacher, K. u. Kiefer, G. (1996): *Kundenzufriedenheit: Analyse, Messbarkeit und Zertifizierung*. Managementwissen für Studium und Praxis. München: Oldenbourg (siehe S. 17).

Scholer u. Turpin 2008
Scholer, F. u. Turpin, A. (2008): „Relevance thresholds in system evaluations“. In: *Proc. 31st SIGIR* (Singapore, 20.–24. Juli 2008). New York: ACM, S. 693–694 (siehe S. 4, 53, 85, 324, 347).

Scholer u. Turpin 2009
Scholer, F. u. Turpin, A. (2009): „Metric and relevance mismatch in retrieval evaluation“. In: *Proc. 5th AIRS* (Sapporo, 21.–23. Okt. 2009). Berlin: Springer, S. 50–62 (siehe S. 61, 85, 116, 144).

Scholer et al. 2008
Scholer, F.; Turpin, A. u. Wu, M. (2008): „Measuring user relevance criteria“. In: *Proc. 2nd EVIA* (Tokyo, 16. Dez. 2008). Tokyo: National Institute of Informatics, S. 47–56 (siehe S. 4, 53, 85, 324, 347).

Schulz et al. 2006
Schulz, N.; Greve, W.; Koch, U. u. Wilmers, N. (2006): „Wie gut erfassen Fragebögen die Qualität der Lehre?“ In: Krampen, G. u. Zayer, H., Krampen, G. u. Zayer, H. (Hrsg.): *Didaktik und Evaluation in der Psychologie*. Göttingen: Hogrefe, S. 75–89 (siehe S. 231).

Sharpsteen u. Bracken 2016
Sharpsteen, C. u. Bracken, C. (2016): *tikzDevice: R graphics output in LaTeX format*. R package version

0.10-1. URL: https://CRAN.R-project.org/package=tikzDevice (verifiziert am 25. 06. 2017) (siehe S. 139).

Shiri u. Revie 2003
Shiri, A. A. u. Revie, C. (2003): „The effects of topic complexity and familiarity on cognitive and physical moves in a thesaurus-enhanced search environment". In: *J. Inf. Sci.* 29 (6), S. 517–526 (siehe S. 28).

Simsek 2011
Simsek, A. (2011): „The relationship between computer anxiety and computer self-efficacy". In: *Contemp. Educ. Technol.* 2 (3), S. 177–187 (siehe S. 36).

Singmann et al. 2016
Singmann, H.; Bolker, B.; Westfall, J. u. Aust, F. (2016): *afex: Analysis of factorial experiments*. R package version 0.16-1. URL: https://CRAN.R-project.org/package=afex (verifiziert am 25. 06. 2017) (siehe S. 139).

Smith 2008
Smith, C. L. (2008): „Searcher adaptation: A response to topic difficulty". In: *Proc. Assoc. Info. Sci. Tech.* 45 (1), S. 1–10 (siehe S. 44, 60).

Smith u. Kantor 2008
Smith, C. L. u. Kantor, P. B. (2008): „User adaptation: Good results from poor systems". In: *Proc. 31st SIGIR* (Singapore, 20.–24. Juli 2008). New York: ACM, S. 147–154 (siehe S. 3, 44, 57, 60 f., 85, 95–100, 102, 107, 111, 114, 116).

Smithson 1994
Smithson, S. (1994): „Information retrieval evaluation in practice: A case study approach". In: *Inf. Process. Manage.* 30 (2), S. 205–221 (siehe S. 54 f., 82, 85, 111 f.).

Smucker u. Jethani 2010a
Smucker, M. D. u. Jethani, C. P. (2010a): „Human performance and retrieval precision revisited". In: *Proc. 33rd SIGIR* (Geneva, 19.–23. Juli 2010). New York: ACM, S. 595–602 (siehe S. 3, 54, 57, 60 f., 85, 93 f., 101, 107, 111, 114, 322).

Smucker u. Jethani 2010b
Smucker, M. D. u. Jethani, C. P. (2010b): „Impact of retrieval precision on perceived difficulty and other user measures". In: *Proc. 4th HCIR* (New Brunswick, 22. Aug. 2010), S. 1–4 (siehe S. 44).

Somers et al. 2003
Somers, T. M.; Nelson, K. u. Karimi, J. (2003): „Confirmatory factor analysis of the end-user computing satisfaction instrument: Replication within an ERP domain". In: *Decis. Sci.* 34 (3), S. 595–621 (siehe S. 119).

Sormunen 2000
Sormunen, E. (2000): „A method for measuring wide range performance of boolean queries in full-text databases". Diss. Tampere: University of Tampere (siehe S. 102, 232, 235).

Sormunen 2002
Sormunen, E. (2002): „Liberal relevance criteria of TREC - counting on negligible documents?" In: *Proc. 25th SIGIR* (Tampere, 11.–15. Aug. 2002). New York: ACM, S. 324–330 (siehe S. 4, 102, 112, 235 f., 324).

Spink u. Greisdorf 2001
Spink, A. u. Greisdorf, H. (2001): „Regions and levels: measuring and mapping users' relevance judgments". In: *J. Am. Soc. Inf. Sci. Technol.* 52 (2), S. 161–173 (siehe S. 102, 111 ff., 324).

Stevens 1975
Stevens, S. S. (1975): *Psychophysics: An introduction to its perceptual, neural and social prospects*. New York: Wiley (siehe S. 112).

Su 1994
Su, L. T. (1994): „The relevance of recall and precision in user evaluation". In: *J. Am. Soc. Inf. Sci.* 45 (3), S. 207–217 (siehe S. 34, 71 f., 75, 81, 86).

Su 1998
Su, L. T. (1998): „Value of search results as a whole as the best single measure of information retrieval performance". In: *Inf. Process. Manage.* 34 (5), S. 557–579 (siehe S. 118).

Su 2003
Su, L. T. (2003): „A comprehensive and systematic model of user evaluation of Web search engines: II.

An evaluation by undergraduates“. In: *J. Am. Soc. Inf. Sci. Technol.* 54 (13), S. 1193–1223 (siehe S. 75, 86, 93).

Szajna u. Scamell 1993
Szajna, B. u. Scamell, R. W. (1993): „The effects of information system user expectations on their performance and perceptions“. In: *Manag. Inf. Syst. Q.* 17 (4), S. 493–525 (siehe S. 38, 72, 83, 86 f., 107 ff., 120, 167, 209, 228, 231).

Tagliacozzo 1977
Tagliacozzo, R. (1977): „Estimating the satisfaction of information users“. In: *Bull. Med. Libr. Assoc.* 65 (2), S. 243–249 (siehe S. 118).

Tang et al. 1999
Tang, R.; Shaw,, William M. u. Vevea, J. L. (1999): „Towards the identification of the optimal number of relevance categories“. In: *J. Am. Soc. Inf. Sci. Technol.* 50 (3), S. 254–264 (siehe S. 102, 112 f., 232).

Tang u. Solomon 1998
Tang, R. u. Solomon, P. (1998): „Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior“. In: *Inf. Process. Manage.* 34 (2–3), S. 237–256 (siehe S. 54, 85).

Tang u. Solomon 2001
Tang, R. u. Solomon, P. (2001): „Use of relevance criteria across stages of document evaluation: On the complementarity of experimental and naturalistic studies“. In: *J. Am. Soc. Inf. Sci. Technol.* 52 (8), S. 676–685 (siehe S. 54 f., 85, 94).

Taylor u. Dalal 2017
Taylor, A. u. Dalal, H. A. (2017): „Gender and information literacy: Evaluation of gender differences in a student survey of information sources“. In: *Coll. Res. Libr.* 78 (1), S. 90–113 (siehe S. 42).

Thomas u. Hawking 2006
Thomas, P. u. Hawking, D. (2006): „Evaluation by comparing result sets in context“. In: *Proc. 15th CIKM* (Arlington, 5.–11. Nov. 2006). New York: ACM, S. 94–101 (siehe S. 62, 75 f., 86).

Thomas et al. 2011
Thomas, P.; Jones, T. u. Hawking, D. (2011): „What deliberately degrading search quality tells us about discount functions“. In: *Proc. 34th SIGIR* (Beijing, 24.–28. Juli 2011). New York: ACM, S. 1107–1108 (siehe S. 97).

Tombros u. Sanderson 1998
Tombros, A. u. Sanderson, M. (1998): „Advantages of query biased summaries in information retrieval“. In: *Proc. 21st SIGIR* (Melbourne, 24.–28. Aug. 1998). New York: ACM, S. 2–10 (siehe S. 95 f., 101, 111).

Toms et al. 2007
Toms, E. G.; O'Brien, H.; Mackenzie, T.; Jordan, C.; Freund, L.; Toze, S.; Dawe, E. u. Macnutt, A. (2007): „Task effects on interactive search: The query factor“. In: *Proc. 6th INEX* (Dagstuhl, 17.–19. Dez. 2007). Bd. 4862. Lecture Notes in Computer Science 4862. Berlin: Springer, S. 359–372 (siehe S. 98).

Toms et al. 2005
Toms, E.; O'Brien, H.; Kopak, R. u. Freund, L. (2005): „Searching for relevance in the relevance of search“. In: *Proc. 5th. CoLIS* (Glasgow, 4.–8. Juni 2005). Bd. 3507. Lecture Notes in Computer Science 3507. Berlin: Springer, S. 59–78 (siehe S. 48, 51, 116).

Torkzadeh u. van Dyke 2002
Torkzadeh, G. u. van Dyke, T. P. (2002): „Effects of training on Internet self-efficacy and computer user attitudes“. In: *Comput. Human Behav.* 18 (5), S. 479–494 (siehe S. 40 f.).

Tsai u. Tsai 2003
Tsai, M.-J. u. Tsai, C.-C. (2003): „Information searching strategies in web-based science learning: The role of internet self-efficacy“. In: *Innov. Educ. Teach. Int.* 40 (1), S. 43–50 (siehe S. 36, 344).

Tse u. Wilton 1988
Tse, D. K. u. Wilton, P. C. (1988): „Models of consumer satisfaction formation: An extension“. In: *J. Mark. Res.* 25 (2), S. 204–212 (siehe S. 19).

Tullis 2007
Tullis, T. S. (2007): „Older adults and the web: Lessons learned from eye-tracking“. In: *Proc. 4th UAHCI* (Beijing, 22.–27. Juli 2007). Berlin: Springer, S. 1030–1039 (siehe S. 38 ff., 91).

Turpin u. Hersh 2001
Turpin, A. H. u. Hersh, W. (2001): „Why batch and user evaluations do not give the same results“. In: *Proc. 24th SIGIR* (New Orleans, 9.–13. Sep. 2001). New York: ACM, S. 225–231 (siehe S. 3, 57–60, 63, 85 f., 97, 114, 162, 321, 335, 339).

Turpin u. Scholer 2006
Turpin, A. u. Scholer, F. (2006): „User performance versus precision measures for simple search tasks“. In: *Proc. 29th SIGIR* (Seattle, 6.–10. Aug. 2006). SIGIR '06. New York: ACM, S. 11–18 (siehe S. 3, 57, 59, 85, 91, 93, 97 ff., 107, 114, 116, 144, 146, 165, 173, 321, 335, 393).

Vakkari 2001
Vakkari, P. (2001): „Changes in search tactics and relevance judgements when preparing a research proposal: A summary of the findings of a longitudinal study“. In: *Inf. Retr.* 4 (3-4), S. 295–310 (siehe S. 54 f., 85).

Vakkari u. Hakala 2000
Vakkari, P. u. Hakala, N. (2000): „Changes in relevance criteria and problem stages in task performance“. In: *J. Doc.* 56 (5), S. 540–562 (siehe S. 23 f., 30, 34, 54 f., 85, 125 f., 325).

Veerasamy u. Belkin 1996
Veerasamy, A. u. Belkin, N. J. (1996): „Evaluation of a tool for visualization of information retrieval results“. In: *Proc. 19th SIGIR* (Zürich, 18.–22. Aug. 1996). New York: ACM, S. 85–92 (siehe S. 114).

Voorhees 2002
Voorhees, E. M. (2002): „The philosophy of information retrieval evaluation“. In: *Proc. 2nd CLEF* (Darmstadt, 3.–4. Sep. 2001). LNCS 2406. Berlin: Springer, S. 355–370 (siehe S. 2).

Wang et al. 2003
Wang, P.; Berry, M. W. u. Yang, Y. (2003): „Mining longitudinal web queries: Trends and patterns“. In: *J. Am. Soc. Inf. Sci.* 54 (8), S. 743–758 (siehe S. 90).

Wang u. Soergel 1998
Wang, P. u. Soergel, D. (1998): „A cognitive model of document use during a research project: Study I. Document selection“. In: *J. Am. Soc. Inf. Sci.* 49 (2), S. 115–133 (siehe S. 51–54, 85, 322).

Wang u. White 1999
Wang, P. u. White, M. D. (1999): „A cognitive model of document use during a research project: Study II. Decisions at the reading and citing stages“. In: *J. Am. Soc. Inf. Sci. Technol.* 50 (2), S. 98–114 (siehe S. 30, 34, 51 f., 54, 85).

Weiner 1985
Weiner, B. (1985): *Human motivation*. New York: Springer (siehe S. 68).

Wenz 2007
Wenz, C. (2007): *JavaScript und AJAX - Das umfassende Handbuch*. Galileo Press. URL: `https://books.google.dk/books?id=4hrV0JnXh4wC` (siehe S. 240).

Werner 2010
Werner, K. (2010): „Größere Zufriedenheit durch bessere Suchmaschinen?: Das Confirmation/Disconfirmation-Paradigma der Kundenzufriedenheit im Kontext des Information Retrieval“. In: *Inform. Wiss. Prax.* 61 (6/7), S. 385–396 (siehe S. 142, 156).

White u. Drucker 2007
White, R. W. u. Drucker, S. M. (2007): „Investigating behavioral variability in web search“. In: *Proc. 16th WWW* (Banff, 8.–12. Mai 2007). New York: ACM, S. 21–30 (siehe S. 20 f., 26, 90).

White u. Dumais 2009
White, R. W. u. Dumais, S. T. (2009): „Characterizing and predicting search engine switching behavior“. In: *Proc. 18th CIKM* (Hong Kong, 2.–6. Nov. 2009). New York: ACM, S. 87–96 (siehe S. 5, 25 f.).

White et al. 2010
White, R. W.; Kapoor, A. u. Dumais, S. T. (2010): „Modeling long-term search engine usage“. In: *Proc. 18th UMAP* (Big Island, 20.–24. Juni 2010). Berlin: Springer, S. 28–39 (siehe S. 26, 82).

White u. Morris 2007
White, R. W. u. Morris, D. (2007): „Investigating the querying and browsing behavior of advanced search engine users“. In: *Proc. 30th SIGIR* (Amsterdam, 23.–27. Juli 2007). New York: ACM, S. 255–262 (siehe S. 27 f., 125).

Whitley 1997
Whitley, B. E. (1997): „Gender differences in computer-related attitudes and behavior: A meta-analysis“. In: *Comput. Human Behav.* 13 (1), S. 1–22 (siehe S. 41).

Wickham 2009
Wickham, H. (2009): *ggplot2: Elegant graphics for data analysis*. New York: Springer. URL: `http://ggplot2.org` (siehe S. 139).

Wilcox 2011a
Wilcox, R. (2011a): *Modern statistics for the social and behavioral sciences: A practical introduction*. Boca Raton: CRC (siehe S. 138).

Wilcox 2011b
Wilcox, R. R. (2011b): *Introduction to robust estimation and hypothesis testing*. Waltham: Academic Press (siehe S. 138).

Wilcox u. Schönbrodt 2014
Wilcox, R. R. u. Schönbrodt, F. D. (2014): *The WRS package for robust statistics in R (version 0.24)*. URL: `https://github.com/nicebread/WRS` (verifiziert am 25. 06. 2017) (siehe S. 138 f.).

Wildemuth 2004
Wildemuth, B. M. (2004): „The effects of domain knowledge on search tactic formulation“. In: *J. Am. Soc. Inf. Sci. Technol.* 55 (3), S. 246–258 (siehe S. 28, 125 f.).

Wildemuth et al. 2014
Wildemuth, B.; Freund, L. u. Toms, E. G. (2014): „Untangling search task complexity and difficulty in the context of interactive information retrieval studies“. In: *J. Doc.* 70 (6), S. 1118–1140 (siehe S. 43, 45, 100).

Wilson 1973
Wilson, P. (1973): „Situational relevance“. In: *Inform. Stor. Retr.* 9 (8), S. 457–471 (siehe S. 49, 329).

Wilson 1999
Wilson, T. D. (1999): „Models in information behaviour research“. In: *J. Doc.* 55 (3), S. 249–270 (siehe S. 8).

Witkin 1973
Witkin, H. A. (1973): „The role of cognitive style in academic performance and in teacher-student relations“. In: *ETS Res. Bull. Ser.* (1), S. 1–58 (siehe S. 20).

Witkin et al. 1977
Witkin, H. A.; Moore, C. A.; Goodenough, D. R. u. Cox, P. W. (1977): „Field-dependent and field-independent cognitive styles and their educational implications“. In: *ETS Res. Bull. Ser.* (2), S. 1–64 (siehe S. 20).

Womser-Hacker 2004
Womser-Hacker, C. (2004): „Theorie des Information Retrieval III: Evaluierung“. In: Kuhlen, R.; Seeger, T. u. Strauch, D., Kuhlen, R.; Seeger, T. u. Strauch, D. (Hrsg.): *Grundlagen der praktischen Information und Dokumentation: Bd. 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis*. 5. Aufl. München: Saur, S. 227–235 (siehe S. 2, 106).

Woodruff et al. 1983
Woodruff, R. B.; Cadotte, E. R. u. Jenkins, R. L. (1983): „Modeling consumer satisfaction processes using experience-based norms“. In: *J. Mark. Res.* 20 (3), S. 296–304 (siehe S. 17).

Xiao u. Dasgupta 2002
Xiao, L. u. Dasgupta, S. (2002): „Measurement of user satisfaction with web-based information systems: An empirical study“. In: *Proc. 8th AMCIS* (Dallas, 9.–11. Aug. 2002). Atlanta: AIS, S. 1149–1155 (siehe S. 119).

Xu u. Mease 2009
Xu, Y. u. Mease, D. (2009): „Evaluating web search using task completion time“. In: *Proc. 32nd SIGIR* (Boston, 19.–23. Juli 2009). New York: ACM, S. 676–677 (siehe S. 78, 87, 96, 116, 335).

Yin et al. 2013
Yin, P.; Luo, P.; Lee, W.-C. u. Wang, M. (2013): „Silence is also evidence: Interpreting dwell time for recommendation from psychological perspective“. In: *Proc. 19th SIGKDD* (Chicago, 11.–14. Aug. 2013). New York: ACM, S. 989–997 (siehe S. 170).

Yuan 1997
Yuan, W. (1997): „End-user searching behavior in information retrieval: A longitudinal study“. In: *J. Am. Soc. Inf. Sci.* 48 (3), S. 218–234 (siehe S. 28).

Zaphiris u. Savtich 2008
Zaphiris, P. u. Savtich, N. (2008): „Age-related differences in browsing the web“. In: *Proc. SPARC: Workshop on promoting independence through new technology* (Reeding, 5. März 2008), S. 1–8 (siehe S. 38 f., 91).

Zeithaml et al. 1993
Zeithaml, V. A.; Berry, L. L. u. Parasuraman, A. (1993): „The nature and determinants of customer expectations of service“. In: *J. Acad. Market. Sci.* 21 (1), S. 1–12 (siehe S. 17, 19, 110).

Zöfel 2003
Zöfel, P. (2003): *Statistik für Psychologen: im Klartext*. München: Pearson (siehe S. 156, 158).

Abbildungsverzeichnis

Tabellenverzeichnis

Variablenübersicht I: Benutzerleistungsmaße

ID	Beschreibung	ID	Beschreibung	ID	Beschreibung
			Dokumentenmengen		
M01	Anz. aufg. Dok.	M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	M33*	Anz. falsch eher rel. bew. rel. Dok.
M02	Anz. aufg. Dok. (erste 10 Dok.)	M18	Anz. richtig rel. bew. Dok. (erste Suche)	M34*	Anz. falsch irrel. bew. Dok.
M03	Anz. aufg. Dok. (erste Suche)	M19	Anz. richtig rel. bew. Dok. (letzte Suche)	M35*	Anz. falsch rel. bew. Dok.
M04	Anz. aufg. Dok. (letzte Suche)	M20*	Anz. aufg. eher irrel. Dok.	M36*	Anz. irrel. bew. Dok.
M05	Anz. aufg. irrel. Dok.	M21*	Anz. aufg. eher rel. Dok.	M37*	Anz. rel. bew. Dok.
M06	Anz. aufg. rel. Dok.	M22*	Anz. aufg. irrel. Dok.	M38*	Anz. rel. bew. Dok. (erste 10 Dok.)
M07	Anz. falsch irrel. bew. Dok.	M23*	Anz. aufg. rel. Dok.	M39*	Anz. rel. bew. Dok. (erste Suche)
M08	Anz. falsch rel. bew. Dok.	M24*	Anz. eher irrel. bew. Dok.	M40*	Anz. rel. bew. Dok. (letzte Suche)
M09	Anz. irrel. bew. Dok.	M25*	Anz. eher rel. bew. Dok.	M41*	Anz. richtig bew. Dok.
M10	Anz. rel. bew. Dok.	M26*	Anz. falsch eher irrel. bew. Dok.	M42*	Anz. richtig eher irrel. bew. Dok.
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	M27*	Anz. falsch eher irrel. bew. eher rel. Dok.	M43*	Anz. richtig eher rel. bew. Dok.
M12	Anz. rel. bew. Dok. (erste Suche)	M28*	Anz. falsch eher irrel. bew. irrel. Dok.	M44*	Anz. richtig irrel. bew. Dok.
M13	Anz. rel. bew. Dok. (letzte Suche)	M29*	Anz. falsch eher irrel. bew. rel. Dok.	M45*	Anz. richtig rel. bew. Dok.
M14	Anz. richtig bew. Dok.	M30*	Anz. falsch eher rel. bew. Dok.	M46*	Anz. richtig rel. bew. Dok. (erste 10 Dok.)
M15	Anz. richtig irrel. bew. Dok.	M31*	Anz. falsch eher rel. bew. eher irrel. Dok.	M47*	Anz. richtig rel. bew. Dok. (erste Suche)
M16	Anz. richtig rel. bew. Dok.	M32*	Anz. falsch eher rel. bew. irrel. Dok.	M48*	Anz. richtig rel. bew. Dok. (letzte Suche)
			Durschnittliche Relevanzbewertung . . .		
B01	irrel. Dok.	B07*	eher irrel. Dok.	B13*	irrel. Dok.
B02	irrel. Dok. (erste Suche)	B08*	eher irrel. Dok. (erste Suche)	B14*	irrel. Dok. (erste Suche)
B03	irrel. Dok. (letzte Suche)	B09*	eher irrel. Dok. (letzte Suche)	B15*	irrel. Dok. (letzte Suche)
B04	rel. Dok.	B10*	eher rel. Dok.	B16*	rel. Dok.
B05	rel. Dok. (erste Suche)	B11*	eher rel. Dok. (erste Suche)	B17*	rel. Dok. (erste Suche)
B06	rel. Dok. (letzte Suche)	B12*	eher rel. Dok. (letzte Suche)	B18*	rel. Dok. (letzte Suche)
			Durschnittliche Betrachtungszeiten . . .		
Z01/-log	aller Dok.	Z11/-log	richtig rel. bew. Dok.	Z21/-log*	irrrel. bew. Dok.
Z02/-log	falsch bew. Dok.	Z12/-log*	eher irrel. bew. Dok.	Z22/-log*	rel. bew. Dok.
Z03/-log	falsch irrel. bew. Dok.	Z13/-log*	eher irrel. Dok.	Z23/-log*	rel. Dok.
Z04/-log	falsch rel. bew. Dok.	Z14/-log*	eher rel. bew. Dok.	Z24/-log*	richtig bew. Dok.
Z05/-log	irrel. bew. Dok.	Z15/-log*	eher rel. Dok.	Z25/-log*	richtig eher irrel. bew. Dok.
Z06/-log	irrel. Dok.	Z16/-log*	falsch eher irrel. bew. Dok.	Z26/-log*	richtig eher rel. bew. Dok.
Z07/-log	rel. bew. Dok.	Z17/-log*	falsch eher rel. bew. Dok.	Z27/-log*	richtig irrel. bew. Dok.
Z08/-log	rel. Dok.	Z18/-log*	falsch irrel. bew. Dok.	Z28/-log*	richtig rel. bew. Dok.
Z09/-log	richtig bew. Dok.	Z19/-log*	falsch rel. bew. Dok.		
Z10/-log	richtig irrel. bew. Dok.	Z20/-log*	irrel. Dok.		
			Sonstige		
S01	Anz. Suchen	S03	Letzte betr. Rankingpos.	S05/-log	Zeit zum ersten richtig rel. bew. Dok.
S02	Erste betr. Rankingpos.	S04	Suchdauer	S06/-log*	Zeit zum ersten richtig rel. bew. Dok.
			Verhältnismaße		
V01	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	V28/PCP	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.	V55	Anz. rel. bew. Dok. / Anz. aufg. Dok.
V02	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	V29	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	V56	Anz. richtig bew. Dok. / Anz. aufg. Dok.
V03	Anz. aufg. rel. Dok. / Anz. rel. Dok. im Korpus	V30	Anz. richtig rel. bew. Dok. / Anz. falsch rel. bew. Dok.	V57	Anz. richtig eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.
V04	Anz. aufg. rel. Dok. / Anz. zurückgeg. rel. Dok.	V31/BP	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.	V58	Anz. richtig eher rel. bew. Dok. / Anz. eher rel. bew. Dok.
V05	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	V32/BR	Anz. richtig rel. bew. Dok. / Anz. rel. Dok. im Korpus	V59	Anz. richtig eher rel. bew. Dok. / Anz. eher rel. Dok. im Korpus
V06	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	V33	Anz. richtig rel. bew. Dok. / Anz. zurückgeg. rel. Dok.	V60	Anz. richtig eher rel. bew. Dok. / Anz. zurückgeg. eher rel. Dok.
V07	Anz. falsch irrel. bew. Dok. / Anz. richtig irrel. bew. Dok.	V34*	Anz. aufg. eher rel. Dok. / Anz. eher rel. Dok. im Korpus	V61*	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.
V08	Anz. falsch rel. bew. Dok. / Anz. aufg. Dok.	V35*	Anz. aufg. eher rel. Dok. / Anz. zurückgeg. eher rel. Dok.	V62*	Anz. richtig irrel. bew. Dok. / Anz. aufg. irrel. Dok.
V09	Anz. falsch rel. bew. Dok. / Anz. rel. bew. Dok.	V36*	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	V63*	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.
V10	Anz. falsch rel. bew. Dok. / Anz. richtig rel. bew. Dok.	V37*	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	V64*	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.
V11	Anz. irrel. bew. Dok. / Anz. aufg. Dok.	V38*	Anz. aufg. rel. Dok. / Anz. rel. Dok. im Korpus	V65*	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. aufg. Dok. (erste 10 Dok.)
V12	Anz. rel. bew. Dok. / Anz. aufg. Dok.	V39*	Anz. aufg. rel. Dok. / Anz. zurückgeg. rel. Dok.	V66*	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)
V13	Anz. richtig bew. Dok. / Anz. aufg. Dok.	V40*	Anz. falsch eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	V67*	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. aufg. Dok. (erste Suche)
V14	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	V41*	Anz. falsch eher irrel. bew. eher rel. Dok. / Anz. eher irrel. bew. Dok.	V68*	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. bew. Dok. (erste Suche)
V15	Anz. richtig irrel. bew. Dok. / Anz. aufg. irrel. Dok.	V42*	Anz. falsch eher irrel. bew. irrel. Dok. / Anz. eher irrel. bew. Dok.	V69*	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. Dok. im Korpus
V16	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.	V43*	Anz. falsch eher irrel. bew. rel. Dok. / Anz. eher irrel. bew. Dok.	V70*	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. zurückgeg. rel. Dok. (erste Suche)
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	V44*	Anz. falsch eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	V71*	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. aufg. Dok. (letzte Suche)
V18	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. aufg. Dok. (erste 10 Dok.)	V45*	Anz. falsch eher rel. bew. eher irrel. Dok. / Anz. eher rel. bew. Dok.	V72*	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. bew. Dok. (letzte Suche)
V19	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)	V46*	Anz. falsch eher rel. bew. irrel. Dok. / Anz. eher rel. bew. Dok.	V73*	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. Dok. im Korpus
V20	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. aufg. Dok. (erste Suche)	V47*	Anz. falsch eher rel. bew. rel. Dok. / Anz. eher rel. bew. Dok.	V74*	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. zurückgeg. rel. Dok. (letzte Suche)
V21	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. bew. Dok. (erste Suche)	V48*	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	V75*	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.
V22	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. Dok. im Korpus	V49*	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	V76*	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.
V23	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. zurückgeg. rel. Dok. (erste Suche)	V50*	Anz. falsch irrel. bew. Dok. / Anz. richtig irrel. bew. Dok.	V77*	Anz. richtig rel. bew. Dok. / Anz. falsch rel. bew. Dok.
V24	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. aufg. Dok. (letzte Suche)	V51*	Anz. falsch rel. bew. Dok. / Anz. aufg. Dok.	V78*	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.
V25	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. bew. Dok. (letzte Suche)	V52*	Anz. falsch rel. bew. Dok. / Anz. rel. bew. Dok.	V79*	Anz. richtig rel. bew. Dok. / Anz. rel. Dok. im Korpus
V26	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. Dok. im Korpus	V53*	Anz. falsch rel. bew. Dok. / Anz. richtig rel. bew. Dok.	V80*	Anz. richtig rel. bew. Dok. / Anz. zurückgeg. rel. Dok.
V27	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. zurückgeg. rel. Dok. (letzte Suche)	V54*	Anz. irrel. bew. Dok. / Anz. aufg. Dok.		

* Leistungsmaß bezieht sich auf vierstufige Relevanzskala in Experiment 3.

Variablenübersicht II: Zufriedenheitsinkikatoren

ID	Beschreibung	Exp. 2	Exp. 3
	Zufriedenheitsitems		
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	x	x
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	x	x
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	x	x
F04	Liefert die Suchmaschine genügend Information?	x	x
F05	Ist die Suchmaschine präzise?	x	x
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	x	x
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	x	x
F08	Ist die Suchmaschine benutzerfreundlich?	x	x
F09	Ist die Suchmaschine einfach zu bedienen?	x	x
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	x	x
F11	Liefert die Suchmaschine aktuelle Information?	x	x
F12	Ist die Suchmaschine erfolgreich?	x	x
F13	Sind Sie mit der Suchmaschine zufrieden?	x	x
F14	Es war einfach, die Aufgabe zu bearbeiten.	x	x
F15	Es war einfach, zu dem Thema zu suchen.	x	
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	x	x
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	x	x
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	x	x
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	x	x
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	x	x
F21	Ich hatte genügend Zeit, um eine effektive Suche durchzuführen.	x	
F22	Ich bin mit den Suchergebnissen zufrieden.	x	x
F23	Ich bin mit meiner Suchleistung zufrieden.	x	x
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	x	x
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	x	x
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	x	x
	Zufriedenheitsskalen		
SK01-M/-F	Zufriedenheit mit der Genauigkeit der Suchmaschine. Umfasst die Items F01, F05 und F06.	x	x
SK02-M/-F	Zufriedenheit mit dem Inhalt der gefundenen Dokumente. Umfasst die Items F02, F03 und F04.	x	x
SK03-M/-F	Zufriedenheit mit der Benutzerfreundlichkeit der Suchmaschine. Umfasst die Items F07 und F08.	x	x
SK04-M/-F	Zufriedenheit mit der Suche. Umfasst die Items F16, F18 und F19.	x	x
SK05-M/-F	Zufriedenheit mit der Aufgabe. Umfasst die Items F14 und F15.	x	
SK06-M/-F	Zufriedenheit mit der eigenen Leistung. Umfasst die Items F20, F21 und F23.	x	
SK07-M/-F	Zufriedenheit mit der Benutzerfreundlichkeit der Suchmaschine. Umfasst die Items F17 und F24.	x	x
SK08-M/-F	Zufriedenheit mit der Suche. Umfasst die Items F01, F02, F03, F04, F05, F16, F17, F18 und F19.	x	x
SK09-M/-F	Zufriedenheit mit der Benutzerfreundlichkeit der Suchmaschine. Umfasst die Items F07, F08 und F24.	x	x
SK10-M/-F	Zufriedenheit mit der Aufgabe. Umfasst die Items F14 und F15.	x	
SK11-M/-F	Zufriedenheit mit der eigenen Leistung. Umfasst die Items F20 und F23.	x	x
SK12-M/-F	Zufriedenheit mit dem Suchergebnis. Umfasst die Items F01, F02, F03, F04, F05, F06 und F07.		x
SK13-M/-F	Zufriedenheit mit der Benutzerfreundlichkeit der Suchmaschine. Umfasst die Items F08 und F09.		x
SK14-M/-F	Zufriedenheit mit der Suche. Umfasst die Items F17, F18 und F19.		x
SK15-M/-F	Zufriedenheit mit der eigenen Leistung. Umfasst die Items F20 und F23.		x
SK16-M/-F	Zufriedenheit mit der Aufgabe. Umfasst die Items F14, F16.		x
SK17-M/-F	Zufriedenheit mit der Suche. Umfasst die Items F01, F03, F04, F05, F07, F18 und F19.		x
SK18-M/-F	Benutzerfreundlichkeit der Suchmaschine. Umfasst die Items F17 und F24.		x
SK19-M/-F	Zufriedenheit mit der eigenen Leistung. Umfasst die Items F20 und F23.		x
SK-A	EUCS-Skala: Zufriedenheit mit der Genauigkeit (Accuracy) der Suchmaschine. Umfasst die Items F05 und F06.	x	x
SK-C	EUCS-Skala: Zufriedenheit mit dem Inhalt (Content) der gefundenen Dokumente. Umfasst die Items F01, F02, F03 und F04.	x	x
SK-E	EUCS-Skala: Zufriedenheit mit der Benutzerfreundlichkeit (Ease of Use) der Suchmaschine. Umfasst die Items F08 und F09.	x	x
SK-T	EUCS-Skala: Zufriedenheit mit der Aktualität (Timeliness) der gefundenen Dokumente. Umfasst die Items F10 und F11.	x	x
SK-K	Außenkriterium des EUCS-Instruments zur Validierung der ermittelten Subskalen. Umfasst die Items F12 und F13.	x	x
SK-E-88	Aus SK-A, SK-C, SK-E und SK-T gebildete Gesamtskala. Umfasst die Items F01, F02, F03, F04, F05, F06, F08, F09, F10 und F11.	x	x
SK-E-09	Aus SK01 bis SK03 gebildete Gesamtskala. Umfasst die Items F01, F02, F03, F04, F05, F06, F07 und F08.	x	x
SK-Z-09	Aus SK04 bis SK07 gebildete Gesamtskala. Umfasst die Items F14, F15, F16, F17, F18, F19, F20, F21, F23 und F24.	x	
SK-G-09	Aus SK08 bis SK11 gebildete Gesamtskala. Umfasst Items F01, F02, F03, F04, F05, F07, F08, F14, F15, F16, F17, F18, F19, F20, F23, F24.	x	
SK-E-13	Aus SK12 und SK13 gebildete Gesamtskala. Umfasst die Items F01, F02, F03, F04, F05, F06, F07, F08 und F09.		x
SK-G-13	Aus SK17 bis SK19 gebildete Gesamtskala. Umfasst die Items F01, F03, F04, F05, F07, F14, F17, F18, F19, F20, F23 und F24.		x
SK-Z-13	Aus SK14 bis SK16 gebildete Gesamtskala. Umfasst die Items F14, F16, F17, F18, F19, F20, F23 und F24.		x
	Erwartungsitems		
E01	Bitte versuchen Sie die Leistung einzuschätzen, die Sie bei der Bearbeitung dieser Aufgabe mit der von Ihnen ausgewählten Suchmaschine erbringen werden.	x	x
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	x	x
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	x	x
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	x	x
E05	Wie wahrscheinlich ist es, dass Sie mit der Suchmaschine sehr zufrieden sind?	x	x
E06-M	Erwartungsskala (E02 , E03, E04, E05)	x	x
E07	Welche Suchmaschine möchten Sie aufgrund Ihrer Erfahrungen bei den vorherigen Aufgaben verwenden, um sich über das Thema zu informieren?	x	

Variablenübersicht III: Kovariaten

ID	Kovariatentyp	Beschreibung	Exp. 2	Exp. 3
K01	Person	Geburtsjahr	x	x
K02	Person	Geschlecht	x	x
K03	Person	Muttersprache	x	x
K04	Erfahrung	Computernutzungsjahre	x	x
K05	Erfahrung	Computernutzungsstunden	x	x
K06	Erfahrung	Domänenwissen	x	
K07	Erfahrung	Selbsteinschätzung Domänenwissen	x	
K08	Erfahrung	Selbsteinschätzung Suchmaschinenwissen	x	x
K09	Erfahrung	Suchmaschinnutzungsjahre	x	x
K10	Erfahrung	Suchmaschinennutzungsstunden	x	x
K11	Erfahrung	Suchmaschinenwissen	x	
K12	Motivation	Ausgangsmotivation	x	

Anhang

A. Verwendete Materialien für Experiment 1

Dieser Anhang enthält Materialien, die im ersten Experiment dieser Arbeit zum Einsatz kommen. Dabei handelt es sich neben den zur Manipulation der Systemleistung verwendeten Rankinglisten des Suchsystems, um die verschiedenen Instruktionstexte zur Manipulation der Erwartungshaltung sowie die verwendeten Fragebogenitems zur Ermittlung von Benutzerzufriedenheit und Sucherfahrung der Testpersonen.

A.1. Verwendete Rankinglisten

In diesem Abschnitt sind die im ersten Experiment verwendeten Rankinglisten für die drei Suchthemen angegeben. Für jedes Suchthema werden zwei unterschiedliche Rankinglisten erzeugt, eine um eine niedrige und eine um eine hohe Retrievalqualität zu simulieren. Insgesamt werden somit für das erste Experiment sechs unterschiedliche Suchergebnislisten benötigt. Zur Generierung der Listen wird dabei auf einen Algorithmus von Turpin und Scholer (2006, S. 14) zurückgegriffen, der für eine vorgegebene Anzahl von relevanten und irrelevanten Dokumenten durch einen randomisierten Austausch der Dokumentenpositionen eine Liste mit festgelegter AvP erzeugt.

A.1.1. Verwendete Rankinglisten für Suchthema 1

Tab. A.1.: Rankingliste des schlechteren Systems für Suchthema 1 (Erneuerbare Energien). Die dargestellte Liste besitzt eine Precision von 0,5 und eine AvP von 0,55. Die insgesamt 100 in dieser Liste enthaltenen Dokumente teilen sich in 50 relevante Dokumente (1) und 50 irrelevante Dokumente (0) auf.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	*1*	*0*	*0*	*0*	*1*	*0*	*1*	*0*	*0*	*1*	*1*	*1*	*0*	*0*	*0*	*1*	*1*	*0*	*1*
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*0*	*1*	*1*	*0*
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
1	*0*	*1*	*1*	*0*	*1*	*1*	*0*	*1*	*0*	*1*	*0*	*1*	*1*	*0*	*0*	*0*	*1*	*0*	*0*
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
0	*0*	*1*	*1*	*1*	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*0*	*0*	*0*
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
0	*1*	*1*	*0*	*1*	*1*	*1*	*0*	*0*	*0*	*1*	*0*	*0*	*1*	*1*	*1*	*0*	*1*	*0*	*0*

Tab. A.2.: Rankinglisten des besseren Systems für Suchthema 1 (Erneuerbare Energien). Die dargestellte Liste besitzt eine Precision von 0,6 und eine AvP von 0,75. Die insgesamt 100 in dieser Liste enthaltenen Dokumente teilen sich in 60 relevante Dokumente (1) und 40 irrelevante Dokumente (0) auf.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	*1*	*0*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*0*	*1*	*1*
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*0*	*1*
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	*1*	*0*	*0*	*0*	*0*	*0*	*1*	*1*	*1*	*0*	*1*	*0*	*0*	*1*	*1*	*1*	*1*	*0*	*1*
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
1	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*1*	*0*	*1*	*1*	*1*	*0*	*0*	*1*	*0*	*0*	*0*
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
0	*1*	*0*	*0*	*0*	*0*	*1*	*0*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*0*

A.1.2. Verwendete Rankinglisten für Suchthema 2

Tab. A.3.: Rankinglisten des schlechteren Systems für Suchthema 2 (Atomtransporte in Deutschland). Die dargestellte Liste besitzt eine Precision von 0,5 und eine AvP von 0,550318. Die insgesamt 96 in dieser Liste enthaltenen Dokumente teilen sich in 48 relevante Dokumente (1) und 48 irrelevante Dokumente (0) auf.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	*1*	*1*	*0*	*0*	*0*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*0*	*1*	*0*	*1*	*1*	*1*
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
0	*0*	*1*	*0*	*0*	*0*	*1*	*0*	*0*	*1*	*1*	*1*	*0*	*0*	*1*	*0*	*0*	*1*	*0*	*0*
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*0*
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
0	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*1*
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
0	*1*	*0*	*1*	*1*	*0*	*1*	*0*	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*0*	-	-	-	-

Tab. A.4.: Rankinglisten des besseren Systems für Suchthema 2 (Atomtransporte in Deutschland). Die dargestellte Liste besitzt eine Precision von 0,6 und eine AvP von 0,75014. Die insgesamt 96 in dieser Liste enthaltenen Dokumente teilen sich in 57 relevante Dokumente (1) und 39 irrelevante Dokumente (0) auf.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*0*
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	*1*	*0*	*1*	*1*	*1*	*0*	*0*	*0*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*0*	*1*	*1*	*1*
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	*0*	*0*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
1	*1*	*1*	*0*	*0*	*1*	*0*	*0*	*0*	*1*	*1*	*0*	*1*	*0*	*1*	*0*	*0*	*0*	*0*	*1*
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
0	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	-	-	-	-

A.1.3. Verwendete Rankinglisten für Suchthema 3

Tab. A.5.: Rankinglisten des schlechteren Systems für Suchthema 3 (Kinderarbeit in Asien). Die dargestellte Liste besitzt eine Precision von 0,5 und eine AvP von 0,550129. Die insgesamt 84 in dieser Liste enthaltenen Dokumente teilen sich in 42 relevante Dokumente (1) und 42 irrelevante Dokumente (0) auf.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	*1*	*0*	*0*	*0*	*1*	*0*	*1*	*1*	*1*	*0*	*1*	*0*	*0*	*1*	*1*	*0*	*1*	*0*	*1*
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	*1*	*0*	*1*	*1*	*0*	*0*	*0*	*0*	*1*	*1*	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*1*
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	*0*	*1*	*1*	*0*	*1*	*1*	*0*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*0*
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
1	*1*	*0*	*0*	*1*	*1*	*1*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*1*
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
1	*0*	*0*	*0*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Tab. A.6.: Rankinglisten des besseren Systems für Suchthema 3 (Kinderarbeit in Asien). Die dargestellte Liste besitzt eine Precision von 0,6 und eine AvP von 0,750398. Die insgesamt 84 in dieser Liste enthaltenen Dokumente teilen sich in 50 relevante Dokumente (1) und 34 irrelevante Dokumente (0) auf.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*0*
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	*0*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*0*	*1*	*0*	*1*
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	*1*	*0*	*0*	*1*	*0*	*1*	*1*	*1*	*0*	*0*	*0*	*1*	*0*	*1*	*0*	*1*	*1*	*0*	*1*
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
1	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*1*	*1*	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*1*	*1*	*0*
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
0	*0*	*1*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

A.2. Testinstruktion

In diesem Abschnitt sind die Instruktionstexte angegeben, die die Probanden auf die Aufgaben des Experiments vorbereiten. Als Erwartungsmanipulation erhalten die Teilnehmer im Rahmen des ersten Experiments je nach Versuchsgruppe unterschiedliche Instruktionen. Zum besseren Vergleich der drei Experimente gliedert sich der vorliegende Abschnitt in die drei Bereiche Erwartungsmanipulation, Einführung und Aufgabenbeschreibung.

A.2.1. Erwartungsmanipulation des ersten Experiments

Niedrige Erwartung: Im Rahmen des Projektseminars Suchmaschinentechnologie soll für die Universität Hildesheim eine neue Suchmaschine für Artikel aus Fachzeitschriften entstehen. Dazu soll die an der Universität Duisburg-Essen von Studenten entwickelte Suchmaschine für Presseartikel Periodikum weiterentwickelt werden. Im Rahmen dieses Benutzertests soll bewertet werden, wie gut diese Suchmaschine in der Lage ist zu einer Suchanfrage relevante Artikel zu liefern und nicht-relevante Artikel zurückzuhalten.

Hohe Erwartung: Die Universität Hildesheim beabsichtigt eine neue Suchmaschine für Artikel aus

Fachzeitschriften anzuschaffen. In die engere Auswahl gekommen ist die Suchmaschine Periodikum von der index Recherche und Suchmaschinentechnologie GmbH. Die Suchmaschinensoftware kostet 20.000 €. Im Rahmen dieses Benutzertests soll bewertet werden, wie gut diese Suchmaschine in der Lage ist zu einer Suchanfrage relevante Artikel zu liefern und nicht-relevante Artikel zurückzuhalten. Die index GmbH hat der Universität Hildesheim zu diesem Zweck eine Demo-Version zur Verfügung gestellt.

A.2.2. Einführung des ersten Experiments

Der Benutzertest wird ca. 30 Minuten dauern. Du bekommst nacheinander drei Suchaufgaben gestellt. Damit alle Teilnehmerinnen die gleichen Voraussetzungen haben, sind die zu verwendenden Suchbegriffe vorgegeben. Deine Aufgabe ist es die Qualität der Ergebnislisten zu bewerten. Im Anschluss folgt ein kurzer Fragebogen. Wenn Du während des Benutzertests irgendwelche Fragen hast, kannst Du diese jederzeit stellen. Unter allen Teilnehmerinnen werden von meinem Fachbereich drei Geldpreise im Wert von 20, 30 und 50€verlost. Alle Untersuchungsdaten werden selbstverständlich anonym ausgewertet und ausschließlich zu wissenschaftlichen Zwecken verwendet. Vielen Dank, dass Du Dich bereit erklärt hast, an diesem Benutzertest teilzunehmen.

A.2.3. Aufgabenbeschreibung des ersten Experiments

Stell Dir für den weiteren Verlauf dieses Benutzertests bitte folgendes Szenario vor: Du bist Journalistin und möchtest Dir für einen Beitrag, den Du demnächst schreiben wirst, einen Überblick über das entsprechende Thema verschaffen. Dazu recherchierst Du mit der Suchmaschine Periodikum nach bereits veröffentlichten Presseartikeln, die das Thema Deines Beitrages betreffen. Zu Demonstrationszwecken befinden sich zurzeit nur Presseartikel aus den Jahren 1994 und 1995 in der Datenbank. Nachdem Du die vorgegebenen Suchbegriffe in das Suchfeld eingegeben und auf *Suche* geklickt hast, erhältst Du eine Ergebnisliste mit Verweisen auf Artikel. Scheint einer dieser Artikel aufgrund der Kurzfassung für Dich relevant zu sein, lässt sich der vollständige Text durch Anklicken des Titels in einem neuen Fenster öffnen. Dort gibt es die Möglichkeit den Artikel als relevant bzw. nicht relevant zu kennzeichnen. Bitte bewerte den Artikel, den Du Dir angesehen hast, bevor Du das Volltext-Fenster wieder schließt. Für jede Suchaufgabe hast Du 10 Minuten Zeit. Wenn Du schon vorher der Meinung bist, Dir einen ausreichenden Überblick über das betreffende Thema verschafft zu haben, kannst Du auch schon vorher mit der nächsten Aufgabe beginnen.

A.3. Items zur Beurteilung der Benutzerzufriedenheit

Dieser Abschnitt enthält eine Auflistung der im ersten Experiment verwendeten Items zur Ermittlung der Benutzerzufriedenheit. Dabei sind alle Items auf einer 7-stufigen Antwortskala von 1 *trifft vollkommen zu* bis 7 *trifft überhaupt nicht zu* zu beantworten.

Tab. A.7.: Items zur Ermittlung der Benutzerzufriedenheit.

Item	**Beschreibung**	**Antwort**		
1	Periodikum entspricht der Vorstellung, die ich von einer Suchmaschine habe.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
2	Periodikum ist einfach zu bedienen.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
3	Ich habe die Recherche mit Periodikum als mühsam und zeitaufwändig empfunden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
4	Ich habe die Recherche mit Periodikum als effizient empfunden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
5	Ich bin mit meinen Rechercheergebnissen zufrieden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
6	Es war schwierig zwischen den Kurzfassungen der einzelnen Artikel auszuwählen.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
7	Ich würde Periodikum jederzeit wieder als Suchmaschine verwenden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
8	Ich bin mit der Qualität der Suchergebnisse zufrieden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
9	Die Artikel hätten besser gefiltert sein können.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
10	Die meisten Artikel waren für die dazugehörigen Suchanfragen relevant.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
11	Die Präsentation der Ergebnisse war übersichtlich.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
12	Die Ergebnislisten waren zu umfangreich.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
13	Die Reihenfolge der Suchergebnisse spiegelte die Relevanz der Artikel wieder.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
14	Die von mir aufgerufenen Artikel waren für die Recherche hilfreich.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
15	Würdest Du den Einsatz von Periodikum als Suchmaschine für Fachzeitschriften in der Hildesheimer Universitätsbibliothek empfehlen?	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu

A.4. Items zur Ermittlung der Sucherfahrung

Um einen möglichen Einfluss der Sucherfahrung bei der statistischen Auswertung berücksichtigen zu können, wird die Sucherfahrung der Testpersonen mit Hilfe von fünf (halb-) offenen Items erfasst. Folgende Items werden den Testpersonen im Zuge dessen im ersten Experiment zur Ermittlung ihrer bisherigen Sucherfahrung vorgelegt:

1. An wie vielen Tagen hast Du in der letzten Woche einen Computer benutzt?
2. Wie viele Stunden verbringst Du in der Woche am Computer?
3. Wie viele Stunden verbringst Du in der Woche im Internet?
4. Welche Suchmaschinen kennst Du?
5. Wie viele verschiedene Suchmaschinen verwendest Du regelmäßig?

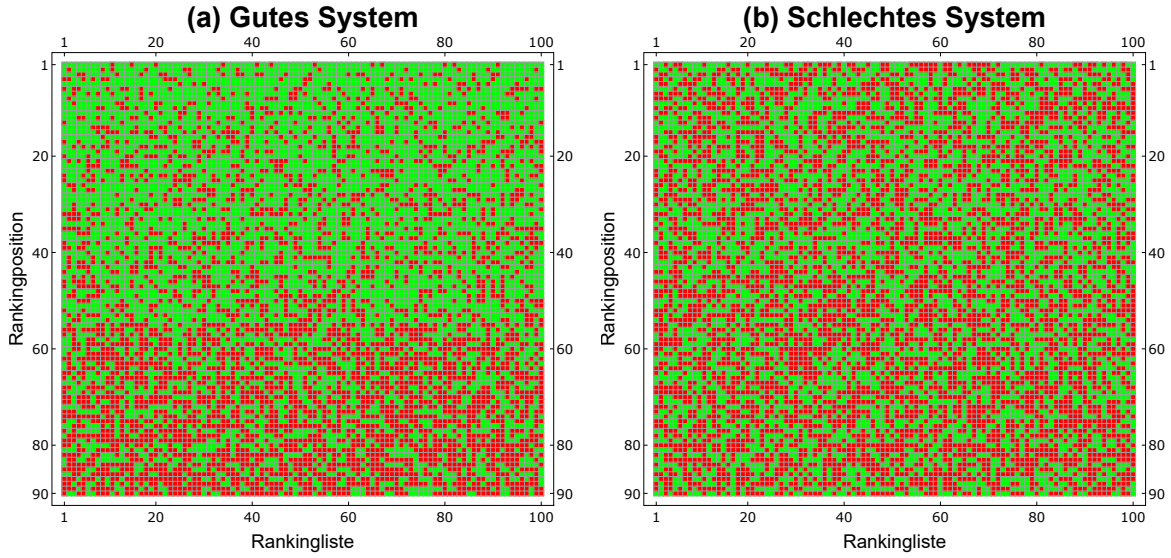
B. Verwendete Materialien für Experiment 2

Dieser Anhang enthält die Materialien, die im zweiten Experiment dieser Arbeit verwendet werden. Dies sind neben den zur Manipulation der Systemleistung zugrunde gelegten Rankinglisten auch die verschiedenen Instruktionstexte zur Manipulation der Erwartungshaltung, die verwendeten Fragebogenitems zur Ermittlung der Benutzerzufriedenheit, der Benutzererwartung und des Domänen- und Suchmaschinenwissens der Testpersonen, eine Übersicht über die eingesetzten Zufriedenheitsskalen, die verwendeten Leistungsmaße zur Beurteilung des Sucherfolgs sowie die im Rahmen der Kovarianzanalyse berücksichtigten Kovariaten.

B.1. Rankinglisten

In diesem Abschnitt sind die im zweiten Experiment verwendeten Rankinglisten für die beiden Systemleistungen angegeben. Für jede Sytemleistung werden 100 unterschiedliche Rankinglisten erzeugt. Insgesamt werden also für das zweite Experiment 200 verschiedene Suchergebnislisten generiert.

Abb. B.1.: Darstellung der in Experiment 2 und 3 verwendeten Rankinglisten für das gute und das schlechte System. Das gute System besitzt eine AvP von 0,75 das schlechte System eine AvP von 0,55. Für jede der 100 verwendeten Listen ist für jedes der 90 enthaltenen Dokumente die Relevanz mit grün (relevant) bzw. rot (irrelevant) markiert. Es ist ergibt sich ein deutlicher Unterschied zwischen den Rankinglisten des guten (a) und schlechten Systems (b).



B.2. Testinstruktion

In diesem Abschnitt sind die Instruktionstexte angegeben, die die Probanden auf die Suchaufgaben des zweiten Experiments vorbereiten. Alle Teilnehmer erhalten im Rahmen dieser Nutzerstudie denselben Instruktionstext, der auch die Erwartungsmanipulation enthält. Zu Beginn der einzelnen Suchaufgaben wird den Probanden dann mitgeteilt, ob sie als nächstes das vermeintlich bessere oder schlechtere System verwenden werden. Zum besseren Vergleich der drei Experimente gliedert sich der vorliegende Abschnitt erneut in die drei Bereiche Einführung,

Erwartungsmanipulation und Aufgabenbeschreibung.

B.2.1. Einführung des zweiten Experiments

Herzlich willkommen zum heutigen Benutzertest und vorab schon einmal vielen Dank, dass Sie sich dazu bereit erklärt haben teilzunehmen, Dieser Test findet im Rahmen des Suchmaschinenprojekts Infofokus statt. In diesem Projekt beschäftigen wir uns mit der Entwicklung effizienterer Algorithmen für Internetsuchmaschinen. In der heutigen Untersuchung sollen zwei Suchmaschinen, die im Zuge der ersten Projektphase entstanden sind, aus einer Benutzerperspektive miteinander verglichen werden. Der Test wird ca. 45 Minuten dauern und besteht im Wesentlichen aus den folgenden drei Teilen: einem Einstiegsfragebogen, einem praktischen Teil, in dem Sie die beiden Suchmaschinen testen und anschließend bewerten werden und einem abschließenden Fragebogenteil. Bevor es jetzt gleich losgeht möchten wir sie noch darauf hinweisen, dass es in dieser Untersuchung nicht um eine Beurteilung Ihrer Person, sondern lediglich um Ihre persönliche Bewertung der beiden Suchmaschinen geht. Alle Ihre Antworten werden selbstverständlich anonym erhoben und ausschließlich zu wissenschaftlichen Zwecken ausgewertet. Bitte beginnen Sie nun mit dem Fragebogen zum thematischen Vorwissen, der vor Ihnen auf dem Tisch liegt.

B.2.2. Erwartungsmanipulation des zweiten Experiments

Im Folgenden werde ich Ihnen den Ablauf des praktischen Teils erläutern. In diesem Abschnitt besteht Ihre Aufgabe darin die beiden im Projekt Infofokus entwickelten Suchmaschinen zu vergleichen. Dazu werden Sie die beiden Suchmaschinen für die Bearbeitung von zwei Suchaufgaben verwenden. Um Verwechselungen zu vermeiden sind beide Suchmaschinen farbig gekennzeichnet. In ersten statistischen Tests hat die blaue Suchmaschine deutlich besser abgeschnitten als die grüne Suchmaschine. Besser heißt in diesem Fall, dass die blaue Suchmaschine im Durchschnitt mehr relevante und gleichzeitig weniger irrelevante Dokumente gefunden hat. Heute möchten wir mit Ihrer Hilfe die Qualität der Suchmaschinen aus Benutzersicht bewerten. Sie bekommen hierzu jeweils eine Suchaufgabe wie zum Beispiel „Suchen Sie Dokumente, die die Nutzung von umweltfreundlicher Energie betreffen“.

B.2.3. Aufgabenbeschreibung des zweiten Experiments

Vom Startbildschirm aus gelangen Sie über diese beiden Buttons zur jeweiligen Suchmaschine. Nachdem Sie Ihre Suchbegriffe in das Suchfeld eingegeben und auf Suche geklickt haben, erhalten Sie eine Ergebnisliste mit Verweisen auf Dokumente. Jede Ergebnisliste besteht aus mehreren Ergebnisseiten zwischen denen Sie über die Seitenauswahl im oberen rechten Inhaltsbereich vor und zurück wechseln können. Selektieren Sie bitte, wie bei einer normalen Suche, nur Links, die Ihrer Ansicht nach auf relevante Dokumente verweisen. Sobald ein Link selektiert wurde, erscheint das jeweilige Webdokument im Inhaltsbereich der Suchmaschine. Um innerhalb des Dokuments zurückzublättern, verwenden Sie bitte anstelle des Zurück-Buttons des Browsers die Blätterfunktion der Suchmaschine. In der linken Navigationsleiste gibt es nun die Möglichkeit dieses Dokument als relevant bzw. irrelevant zu kennzeichnen. Bitte bewerten Sie das Webdokument, das Sie sich angesehen haben, bevor Sie über den Zurück-Button wieder zur Ergebnisliste gelangen. Bevor es nun tatsächlich losgeht, noch ein kurzer Hinweis. Alle Untersuchungsteilnehmer haben die Chance einen von drei Geldpreisen zu 20, 30 oder 50 € zu gewinnen. Sieger ist, wer zu einer Suchaufgabe die meisten relevanten und die wenigsten irrelevanten Dokumente findet. Und nun viel Spaß bei der ersten Aufgabe. Den Aufgabenzettel erhalten Sie von Ihrem Testleiter.

B.3. Items zur Beurteilung der Benutzerzufriedenheit

Dieser Abschnitt enthält eine Auflistung der im zweiten Experiment verwendeten Items zur Ermittlung der Zufriedenheit der Testpersonen. Während die 13 aus dem EUCS-Instrument über-

nommenen Items auf einer 5-stufigen Antwortskala (*fast nie - manchmal - in der Hälfte der Fälle - meistens - fast immer*) zu beantworten sind, werden die 13 Zusatzitems anhand einer 7-stufigen Antwortskala (*trifft überhaupt nicht zu - trifft nicht zu - trifft eher nicht zu - weder noch - trifft eher zu - trifft zu - trifft vollkommen zu*) erfasst.

Tab. B.1.: Items zur Beurteilung der Benutzerzufriedenheit mit 5-stufiger Skala. Die dargestellten Items stammen aus dem EUCS-Instrument von Doll und Torkzadeh (1988).

Item	**Beschreibung**	**Antwort**		
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	fast nie	1 2 3 4 5	fast immer
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	fast nie	1 2 3 4 5	fast immer
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	fast nie	1 2 3 4 5	fast immer
F04	Liefert die Suchmaschine genügend Information?	fast nie	1 2 3 4 5	fast immer
F05	Ist die Suchmaschine präzise?	fast nie	1 2 3 4 5	fast immer
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	fast nie	1 2 3 4 5	fast immer
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	fast nie	1 2 3 4 5	fast immer
F08	Ist die Suchmaschine benutzerfreundlich?	fast nie	1 2 3 4 5	fast immer
F09	Ist die Suchmaschine einfach zu bedienen?	fast nie	1 2 3 4 5	fast immer
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	fast nie	1 2 3 4 5	fast immer
F11	Liefert die Suchmaschine aktuelle Information?	fast nie	1 2 3 4 5	fast immer
F12	Ist die Suchmaschine erfolgreich?	fast nie	1 2 3 4 5	fast immer
F13	Sind Sie mit der Suchmaschine zufrieden?	fast nie	1 2 3 4 5	fast immer

Tab. B.2.: Items zur Beurteilung der Benutzerzufriedenheit mit 7-stufiger Skala.

Item	Beschreibung	Antwort		
F14	Es war einfach, die Aufgabe zu bearbeiten.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F15	Es war einfach, zu dem Thema zu suchen.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F21	Ich hatte genügend Zeit, um eine effektive Suche durchzuführen.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F22	Ich bin mit den Suchergebnissen zufrieden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F23	Ich bin mit meiner Suchleistung zufrieden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	trifft vollkommen zu	1 2 3 4 5 6 7	trifft überhaupt nicht zu

B.4. Items zur Beurteilung der Benutzererwartungen

Um die aus der Systemnutzung resultierende Einstellung der Testpersonen zu den beiden getesteten Suchmaschinen erheben zu können, werden die Teilnehmer im Anschluss an die zweite Aufgabe gebeten sich vorzustellen, zwei weitere fiktive Suchaufgaben zu bearbeiten und jeweils sechs erwartungsbezogene Frageitems zu beantworten. Die der Ermittlung der Benutzererwartungen zugrunde gelegten Aufgabenstellungen lauten wie folgt:

1. Stellen Sie sich vor, Sie müssen eine Hausarbeit über die Nachhaltigkeitsverordnung des Erneuerbare-Energien-Gesetz schreiben.

2. Stellen Sie sich vor, Sie planen ein Praktikum im Bereich erneuerbare Energien zu machen.

Tabelle B.4 umfasst die zur Ermittlung der Benutzererwartungen verwendeten Frageitems, die im Anschluss an jede der beiden fiktiven Suchaufgaben erhoben werden. Bis auf die ersten beiden Fragen sind alle Items auf einer 7-stufigen Antwortskala (*sehr unwahrscheinlich - unwahrscheinlich - ziemlich unwahrscheinlich - neutral - ziemlich wahrscheinlich - wahrscheinlich - sehr wahrscheinlich*) zu beantworten.

Tab. B.4.: Items zur Beurteilung der Benutzererwartungen. Die Bezeichnung der Items entspricht ihrer Nummerierung in den entsprechenden Ergebniskapiteln, die Sortierung in dieser Tabelle der Reihenfolge im Fragebogen. Bei E06-M handelt es sich um eine aus den Items E02 bis E05 gebildete Mittelwertskala, die somit nicht im Testfragebogen enthalten ist.

Item	**Beschreibung**	**Antwort**		
E07	Welche Suchmaschine möchten Sie aufgrund Ihrer Erfahrungen bei den vorherigen Aufgaben verwenden, um sich über das Thema zu informieren?	◯ Die blaue Suchmaschine ◯ Die grüne Suchmaschine		
E01	Bitte versuchen Sie die Leistung einzuschätzen, die Sie bei der Bearbeitung dieser Aufgabe mit der von Ihnen ausgewählten Suchmaschine erbringen werden.	Ich glaube, ich werde in zehn Minuten ... relevante Dokumente mit der von mir ausgewählten Suchmaschine finden.		
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	sehr unwahrscheinlich	1 2 3 4 5 6 7	sehr wahrscheinlich
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	sehr unwahrscheinlich	1 2 3 4 5 6 7	sehr wahrscheinlich
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	sehr unwahrscheinlich	1 2 3 4 5 6 7	sehr wahrscheinlich
E05	Wie wahrscheinlich ist es, dass Sie mit der Suchmaschine sehr zufrieden sind?	sehr unwahrscheinlich	1 2 3 4 5 6 7	sehr wahrscheinlich
E06-M	Mittelwertsskala über die Items E02 bis E05			

B.5. Items zur Beurteilung des Domänenwissens

In diesem Abschnitt sind die zur Ermittlung des Domänenwissens der Testteilnehmer entwickelten Frageitems wiedergegeben. Die Probanden werden gebeten, jeweils diejenige Antwort auszusuchen und anzukreuzen, die Ihrer Meinung nach zutrifft. Für den Fall, dass sie die Frage nicht beantworten können, werden sie gebeten, nicht zu raten, sondern stattdessen das Kästchen *weiß ich nicht* anzukreuzen.

Tab. B.5.: Items zur Beurteilung des Domänenwissens. Die jeweils zutreffende Antwort ist in der Tabelle fett hervorgehoben.

Item	Beschreibung	Antwort
1	Wie werden erneuerbare Energien noch genannt?	**(a) Regenerative Energien** (b) Kinetische Energien (c) Haltbare Energien (d) Wiederherstellbare Energien (e) weiß ich nicht
2	Was meint das Wort Geothermie?	(a) Die Aufheizung der erdnahen Luftschichten durch Sonnenstrahlung. (b) Die allmähliche Abkühlung der Erde. **(c) Die Nutzung der Wärme des Erdinneren.** (d) Die aktive Kühlung in äquatornahen Regionen. (e) weiß ich nicht
3	Was sind Offshore-Windparks?	(a) Windenergieanlagen, die steuerbegünstigten Strom liefern. (b) Windenergieanlagen auf der Küste vorgelagerten Inseln. **(c) Windenergieanlagen auf dem offenen Meer.** (d) Windenergieanlagen, die staatlich subventioniert werden. (e) weiß ich nicht
4	Sonnenenergie wird in zwei verschiedenen Formen genutzt. Wie heißt die Variante, bei der Licht in elektrischen Strom umgewandelt wird?	(a) Kinesiologie **(b) Photovoltaik** (c) Plasmasynthese (d) Photosynthese (e) weiß ich nicht
5	Was macht ein Gezeitenkraftwerk?	**(a) Nutzt die Energie aus dem ständigen Wechsel von Ebbe und Flut.** (b) Nutzt die Energie der kontinuierlichen Meereswellen. (c) Nutzt die kinetische Energie von Meeresströmungen. (d) Nutzt den Unterschied im Salzgehalt zwischen Meerwasser und Süßwasser. (e) weiß ich nicht
6	Aus welcher Pflanze wird schon seit mehreren Jahren Biodiesel hergestellt?	(a) Kohl **(b) Raps** (c) Weizen (d) Mais (e) weiß ich nicht
7	Mit Sonnenkollektoren kann man?	(a) Aus Sonnenlicht Strom erzeugen. (b) Tageszeit nach dem Stand der Sonne bestimmen. (c) Wetteränderungen sattelitengestützt vorhersagen. **(d) Aus Sonnenlicht Warmwasser und Heizenergie erzeugen.** (e) weiß ich nicht
8	Was ist das wichtigste Material einer Solarzelle?	(a) Glas **(b) Silizium** (c) Aluminium (d) Quecksilber (e) weiß ich nicht

B.6. Items zur Beurteilung des Suchmaschinenwissens

In diesem Abschnitt sind die zur Ermittlung des Suchmaschinenwissens der Testteilnehmer entwickelten Frageitems wiedergegeben. In Tabelle B.6 sind die fünf Items zur Erfassung des theoretischen Suchmaschinenwissens dargestellt. Die Probanden werden gebeten, jeweils diejenige Alternative auszusuchen und anzukreuzen, die ihrer Ansicht nach den jeweiligen Begriff am besten charakterisiert. Wie im Fall des Domänenwissens, werden die Probanden gebeten, nicht zu raten, sondern das Kästchen *weiß ich nicht* ankreuzen, falls ihnen der betreffende Begriff nicht geläufig ist. In Tabelle B.7 sind die sieben Items zur Erfassung des praktischen Suchmaschinenwissens zusammengestellt. Auch hier werden die Probanden gebeten, für jede geschilderte Situation, diejenige Antwort auszusuchen und anzukreuzen, die Ihrer Einschätzung nach am ehesten zutrifft und im Zweifel die *weiß ich nicht* Antwortoption auszuwählen.

Tab. B.6.: Items zur Beurteilung des theoretischen Suchmaschinenwissens. Die jeweils zutreffende Antwort ist in der Tabelle fett hervorgehoben.

Item	Begriff	Antwort
1	Volltextsuchmaschine	**(a) Suchmaschine, die auf einer automatischen Indexierung von Webdokumenten basiert.** (b) Suchmaschine zum Auffinden von Volltexten. (c) Spezielle Suchmaschine für Journalisten, die nach aktuellen Pressemitteilungen sucht. (d) Alternative Bezeichnung für eine Plagiatsuchmaschine. (e) weiß ich nicht
2	Webkatalog	(a) Katalog, der alle Internetseiten auflistet und per Post zugestellt wird. **(b) Suchmaschine, bei der die Daten manuell in eine bestimmte Ordnung gebracht werden.** (c) Reiseprospekte, die man downloaden kann. (d) Suchmaschine speziell für Jobs in der IT-Branche. (e) weiß ich nicht
3	Metasuchmaschine	(a) Spezielle Suchmaschine für medizinische Fortbildungsveranstaltungen. **(b) Suchmaschine, die mehrere Suchmaschinen gleichzeitig durchsucht.** (c) Personalisierte Suchmaschine, die von Nutzern selbst konzipiert wurden. (d) Oberbegriff für Suchmaschinen, die sich auf einen bestimmten Themenbereich oder eine geografische Region beziehen. (e) weiß ich nicht
4	Stoppworte	(a) Ethisch verwerfliche Worte. **(b) Worte wie die oder für, mit denen man nicht auf Inhalte schließen kann.** (c) Worte, die normalerweise dazu führen, dass sich Suchalgorithmen in Schleifen verfangen. (d) Synonyme für Stopp. (e) weiß ich nicht
5	PageRank	PageRank ist ein Verfahren zur Bewertung der Wichtigkeit von Webseiten, bei dem davon ausgegangen wird, dass ... (a) eine Webseite umso relevanter ist, je mehr Einzelseiten sie hat. (b) eine Webseite umso relevanter ist, je mehr externe Links auf den Seiten zu finden sind. (c) eine Webseite umso relevanter ist, je mehr interne Links auf den Seiten zu finden sind. **(d) eine Webseite umso relevanter ist, je mehr andere Homepages auf sie verweisen.** (e) weiß ich nicht

Tab. B.7.: Items zur Beurteilung des praktischen Suchmaschinenwissens. Die jeweils zutreffende Antwort ist in der Tabelle fett hervorgehoben.

Item	Beschreibung	Antwort
1	Welche der folgenden Suchanfragen findet (potentiell) eine größere Anzahl von Dokumenten?	(a) Universität Hildesheim **(b) Universität NOT Hildesheim** (c) weiß ich nicht
2	Welche der folgenden Suchanfragen findet (potentiell) eine größere Anzahl von Dokumenten?	(a) „erneuerbare Energien“ **(b) erneuerbare Energien** (c) weiß ich nicht
3	Welche der folgenden Suchanfragen findet (potentiell) eine größere Anzahl von Dokumenten?	(a) Bank AND Kreditinstitut AND Geldinstitut AND Kreditanstalt AND Sparkasse **(b) Bank OR Kreditinstitut OR Geldinstitut OR Kreditanstalt OR Sparkasse** (c) weiß ich nicht
4	Welche der folgenden Suchanfragen findet (potentiell) eine größere Anzahl von Dokumenten?	(a) Auto **(b) Auto*** (c) weiß ich nicht

Fortsetzung auf nächster Seite

Tab. B.7 (Fortsetzung)

Item	Beschreibung	Antwort
5	Nehmen Sie an, bei einer Suchmaschine steht das folgende Dokument zur Verfügung: Wir suchen: Verstärkung für unser hochmotiviertes Team. Sie sind teamfähig, kreativ, ausdauernd, geduldig, gutaussehend, sportlich, innovativ, intelligent, mit Führungsqualitäten, Hochschulabschluss inkl. Doktortitel in drei Disziplinen und bereit unkompensierte Überstunden zu arbeiten und auf Ihre Ferien zu verzichten. Wir bieten firmeneigene Parkplätze zu vergünstigten Konditionen sowie unseren traditionellen Osterausflug für alle MitarbeiterInnen. Sie suchen nach Stelleninseraten mit der Suchanfrage *Stelleninserate*. Würden Sie mit dieser Anfrage das obige Dokument finden?	(a) Ja, das Dokument wird gefunden. **(b) Nein, das Dokument wird nicht gefunden.** (c) weiß ich nicht
6	Auf die Suchanfrage *Peter Müller* erhalten Sie die folgenden Dokumente A und B. Welches wird in der Trefferliste wahrscheinlich als erstes angezeigt? A: Scientology stampft Werbung mit Peter Müller ein. Die Scientology Kirche hat sich verpflichtet, ihre Werbung für das Buch „Dianetik“mit dem Skistar Peter Müller einzustampfen. Der Ex-Skirennfahrer wurde von Scientology ohne das Einverständnis des für die Werbung mit Skistars zuständigen Swiss Ski Pool eingespannt. Wie der Pool in einer Mitteilung vom Freitag weiter schreibt, hat sich Scientology wegen der Androhung prozessualer Schritte verpflichtet, die Persönlichkeitsrechte des Betroffenen vollumfänglich zu respektieren. B: Europarat-Parlamentarier verurteilen iranisches Regime. Über 110 Abgeordnete der Parlamentarischen Versammlung des Europarates, darunter zahlreiche Schweizer, haben in einer am Freitag veröffentlichten Erklärung das iranische Regime verurteilt. Sie beschuldigten Teheran, seine Unterdrückungspolitik fortzusetzen und politische Häftlinge zu massakrieren sowie den internationalen Terrorismus zu unterstützen. Die Erklärung ruft überdies zur Anerkennung des von Massud Radschawi geleiteten „Nationalrats des iranischen Widerstandes“auf. Von den Schweizer Europarat-Parlamentariern unterzeichneten den Aufruf Doris Morf, Michel Flückiger, Peter Sager, Massimo Pini, Bernhard Seiler, Andreas Müller, Fulvio Caccia und Hans Jörg Huber.	**(a) Dokument A** (b) Dokument B (c) weiß ich nicht

Fortsetzung auf nächster Seite

Tab. B.7 (Fortsetzung)

Item	Beschreibung	Antwort
7	Auf die Suchanfrage *Waldsterben* erhalten Sie die folgenden Dokumente A und B. Welches wird in der Trefferliste wahrscheinlich als erstes angezeigt? A: Saurer Regen zerstört auch das Leben im Meer. Der saure Regen ist nicht nur mitverantwortlich für das Waldsterben und für die Schädigung von Bächen und Seen. Offenbar werden auch Pflanzen und Tiere an der amerikanischen Atlantikküste durch die sauren Niederschläge schwer in Mitleidenschaft gezogen. Dies ist das Ergebnis einer Studie des US- Umweltschutzfonds, die am Dienstag in Washington veröffentlicht wurde. Sauerstoffmangel führt an der Atlantikküste in bestimmten Abständen zum Absterben von Wasserpflanzen, Krustentieren und anderen Formen marinen Lebens. Nach einer kürzlich in den USA veröffentlichten Studie ist an diesem Erstickungstod stärker als bisher angenommen der saure Regen schuld. B: Waldsterben für die Schweiz keine existentielle Bedrohung. Rodolph Schläpfer, Direktor der Eidg. Anstalt für Forstliches Versuchswesen (EAFV), ist der Ansicht, das Waldsterben bedeute für die Schweiz keine existentielle Gefährdung. Schläpfer relativierte mit dieser Aussage in einem Interview mit der Schweizer Illustrierten die früheren Devisen der EAFV. Laut Schläpfer sprachen vor drei Jahren noch gute Gründe für die Annahme, das Waldsterben sei eine lebensgefährliche Bedrohung für die Schweiz. Die Entwicklung sei aber nicht so dramatisch verlaufen, wie früher angenommen. Nach seiner Ansicht sollte der Begriff Baumkrankheit neu definiert werden. Gegenwärtig wird ein Baum als krank eingestuft, wenn ein Nadel- oder Laubverlust von 10 Prozent vorliegt. Schläpfer ist der Ansicht, es wäre möglicherweise besser, zur Definition der Baumkrankheit einen Schädigungsanteil von 25 Prozent zu verwenden. Der Direktor der EAFV wandte sich im Weiteren auch gegen eine wörtliche Interpretation des Begriffes „Waldsterben“. Er sagte, er betrachte das Waldsterben als generellen Vitalitätsverlust des Waldes. Dieser müsse aber nicht unbedingt zum Sterben führen.	(a) Dokument A **(b) Dokument B** (c) weiß ich nicht

B.7. Items zur Beurteilung der Sucherfahrung

Um wie im ersten Experiment einen möglichen Einfluss der Sucherfahrung bei der statistischen Auswertung berücksichtigen zu können, wird darüber hinaus die Sucherfahrung der Testpersonen mit Hilfe von sechs (halb-) offenen Items erfasst. Folgende Items werden den Testpersonen im Zuge dessen im zweiten Experiment zur Ermittlung ihrer bisherigen Sucherfahrung vorgelegt:

1. Seit wie vielen Jahren nutzen Sie bereits einen Computer?
2. Wie viel Zeit in Stunden verbringen Sie durchschnittlich pro Woche mit dem Computer?
3. Seit wie vielen Jahren nutzen Sie bereits Suchmaschinen?
4. Wie viel Zeit in Stunden verbringen Sie durchschnittlich pro Woche mit Suchmaschinen?
5. Welche Suchmaschinen nutzen Sie?
6. Kennen Sie weitere Suchmaschinen? Wenn ja, welche?

B.8. Skalen zur Beurteilung der Benutzerzufriedenheit

Dieser Abschnitt bietet eine Übersicht über die im zweiten Experiment ermittelten Zufriedenheitsskalen. Die empirische Herleitung dieser Skalen ist ausführlich in Abschnitt 6.4.4.2 beschrieben. Neben den im Rahmen der Faktorenanalyse ermittelten Skalen gibt Tabelle B.8 auch

Aufschluss über die im Zuge der Auswertung gebildeten Gesamtskalen (SK-E-88, SK-E-09, SK-Z-09 u. SK-G-09).

Tab. B.8.: Übersicht über verwendete Benutzerzufriedenheitsskalen. Die Buchstaben M bzw. F am Ende der Skalenbezeichnung geben an, ob zu ihrer Berechnung der Mittelwert der Items (M) oder der entsprechende Faktorwert (F) zugrunde gelegt wird.

Skala	Beschreibung	Items
SK01-M/SK01-F	Genauigkeit	F02, F03, F04
SK02-M/SK02-F	Inhalt	F01, F05, F06
SK03-M/SK03-F	Benutzerfreundlichkeit	F07, F08
SK04-M/SK04-F	Suche	F16, F18, F19
SK05-M/SK05-F[a]	Aufgabe	F14, F15
SK06-M/SK06-F	Eigenleistung	F20, F21, F23
SK07-M/SK07-F	Benutzerfreundlichkeit	F17, F24
SK08-M/SK08-F	Suche	F01, F02, F03, F04, F05, F16, F17, F18, F19
SK09-M/SK09-F	Benutzerfreundlichkeit	F07, F08, F24
SK10-M/SK10-F[a]	Aufgabe	F14, F15
SK11-M/SK11-F	Eigenleistung	F20, F23
SK-A	Accuracy (EUCS)	F05, F06
SK-C	Content (EUCS)	F01, F02, F03, F04
SK-E	Ease of Use (EUCS)	F08, F09
SK-T	Timeliness (EUCS)	F10, F11
SK-K	Kriteriumsskala	F12, F13
SK-E-88	EUCS-Skala-1988	F01, F02, F03, F04, F05, F06, F08, F09, F10, F11
SK-E-09	EUCS-Skala-2009	F01, F02, F03, F04, F05, F06, F07, F08
SK-Z-09	Zusatzskala-2009	F14, F15, F16, F17, F18, F19, F20, F21, F23, F24
SK-G-09	Gesamtskala-2009	F01, F02, F03, F04, F05, F07, F08, F14, F15, F16, F17, F18, F19, F20, F23, F24

[a] Die Mittelwertsskalen SK05-M und SK10-M sind identisch.

B.9. Maße zur Beurteilung der Benutzerleistung

Dieser Abschnitt bietet eine Übersicht über die im zweiten Experiment verwendeten Benutzerleistungsmaße. Entsprechend ihrer Einführung im Rahmen der Operationalisierung der abhängigen Variablen (vgl. Abschn. 6.3.2) sind die in den Tabellen B.9 bis B.13 beschriebenen Benutzerleistungsmaße in die folgenden fünf Variablengruppen untergliedert: Dokumentenmengen, Bewertungen von Dokumentenmengen, durchschnittliche Betrachtungszeiten unterschiedlicher Dokumentenmengen, Verhältnisse von Dokumentenmengen und sonstige Leistungsmaße.

Tab. B.9.: Dokumentenmengen bei binärer Relevanzskala (Variablengrp. 1). Bei der Zählung der Dokumentenmengen wird jedes Dokument nur einmal berücksichtigt. Als Dokumentenbewertung geht jeweils die letzte Bewertung eines Dokuments über alle Suchanfragen hinweg in die Betrachtung ein.

ID	Beschreibung	Kurzform
M01	Anzahl aufgerufener Dokumente	Anz. aufg. Dok.
M02	Anzahl aufgerufener Dokumente der ersten zehn angezeigten Dokumente	Anz. aufg. Dok. (erste 10 Dok.)
M03	Anzahl aufgerufener Dokumente der ersten durchgeführten Suche	Anz. aufg. Dok. (erste Suche)
M04	Anzahl aufgerufener Dokumente der letzten durchgeführten Suche	Anz. aufg. Dok. (letzte Suche)
M05	Anzahl aufgerufener irrelevanter Dokumente	Anz. aufg. irrel. Dok.
M06	Anzahl aufgerufener relevanter Dokumente	Anz. aufg. rel. Dok.
M07	Anzahl falsch irrelevant bewerteter Dokumente	Anz. falsch irrel. bew. Dok.
M08	Anzahl falsch relevant bewerteter Dokumente	Anz. falsch rel. bew. Dok.
M09	Anzahl irrelevant bewerteter Dokumente	Anz. irrel. bew. Dok.
M10	Anzahl relevant bewerteter Dokumente	Anz. rel. bew. Dok.

Fortsetzung auf nächster Seite

Tab. B.9 (Fortsetzung) Variablengrp. 1

ID	Beschreibung	Kurzform
M11	Anzahl relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente	Anz. rel. bew. Dok. (erste 10 Dok.)
M12	Anzahl relevant bewerteter Dokumente der ersten durchgeführten Suche	Anz. rel. bew. Dok. (erste Suche)
M13	Anzahl relevant bewerteter Dokumente der letzten durchgeführten Suche	Anz. rel. bew. Dok. (letzte Suche)
M14	Anzahl richtig bewerteter Dokumente	Anz. richtig bew. Dok.
M15	Anzahl richtig irrelevant bewerteter Dokumente	Anz. richtig irrel. bew. Dok.
M16[a]	Anzahl richtig relevant bewerteter Dokumente	Anz. richtig rel. bew. Dok.
M17	Anzahl richtig relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente	Anz. richtig rel. bew. Dok. (erste 10 Dok.)
M18	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche	Anz. richtig rel. bew. Dok. (erste Suche)
M19	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche	Anz. richtig rel. bew. Dok. (letzte Suche)

[a] M16 entspricht dem im ersten Experiment verwendeten Benutzerleistungsmaß RRD.

Tab. B.10.: Bewertungen von Dokumentenmengen bei binärer Relevanzskala (Variablengrp. 2). Als Dokumentenbewertung geht jeweils die letzte Bewertung eines Dokuments über alle Suchanfragen in die Betrachtung ein. Zur Bestimmung der mittleren Bewertung werden als relevant bewertete Dokumente mit 1, irrelevant bewertete Dokumente mit 0 kodiert.

ID	Beschreibung	Kurzform
B01	Durchschnittliche Bewertung irrelevanter Dokumente	Durchschn. Bew. irrel. Dok.
B02	Durchschnittliche Bewertung irrelevanter Dokumente der ersten durchgeführten Suche	Durchschn. Bew. irrel. Dok. (erste Suche)
B03	Durchschnittliche Bewertung irrelevanter Dokumente der letzten durchgeführten Suche	Durchschn. Bew. irrel. Dok. (letzte Suche)
B04	Durchschnittliche Bewertung relevanter Dokumente	Durchschn. Bew. rel. Dok.
B05	Durchschnittliche Bewertung relevanter Dokumente der ersten durchgeführten Suche	Durchschn. Bew. rel. Dok. (erste Suche)
B06	Durchschnittliche Bewertung relevanter Dokumente der letzten durchgeführten Suche	Durchschn. Bew. rel. Dok. (letzte Suche)

Tab. B.11.: Durchschnittliche Betrachtungszeiten unterschiedlicher Dokumentenmengen bei binärer Relevanzskala (Variablengrp. 3). Berücksichtigt wird jeweils der erste Aufruf eines Dokuments.

ID	Beschreibung	Kurzform
Z01/Z01-log	Durchschnittliche Betrachtungszeit aller Dokumente	Durchschn. Betrachtungsz. aller Dok.
Z02/Z02-log	Durchschnittliche Betrachtungszeit falsch bewerteter Dokumente	Durchschn. Betrachtungsz. falsch bew. Dok.
Z03/Z03-log	Durchschnittliche Betrachtungszeit falsch irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. falsch irrel. bew. Dok.
Z04/Z04-log	Durchschnittliche Betrachtungszeit falsch relevant bewerteter Dokumente	Durchschn. Betrachtungsz. falsch rel. bew. Dok.

Fortsetzung auf nächster Seite

Tab. B.11 (Fortsetzung) Variablengrp. 3

ID	Beschreibung	Kurzform
Z05/Z05-log	Durchschnittliche Betrachtungszeit irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. irrel. bew. Dok.
Z06/Z06-log	Durchschnittliche Betrachtungszeit irrelevanter Dokumente	Durchschn. Betrachtungsz. irrel. Dok.
Z07/Z07-log	Durchschnittliche Betrachtungszeit relevant bewerteter Dokumente	Durchschn. Betrachtungsz. rel. bew. Dok.
Z08/Z08-log	Durchschnittliche Betrachtungszeit relevanter Dokumente	Durchschn. Betrachtungsz. rel. Dok.
Z09/Z09-log	Durchschnittliche Betrachtungszeit richtig bewerteter Dokumente	Durchschn. Betrachtungsz. richtig bew. Dok.
Z10/Z10-log	Durchschnittliche Betrachtungszeit richtig irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. richtig irrel. bew. Dok.
Z11/Z11-log	Durchschnittliche Betrachtungszeit richtig relevant bewerteter Dokumente	Durchschn. Betrachtungsz. richtig rel. bew. Dok.

Tab. B.12.: Verhältnisse von Dokumentenmengen bei binärer Relevanzskala (Variablengrp. 4).

ID	Beschreibung	Kurzform	Formel
V01	Anzahl aufgerufener irrelevanter Dokumente / Anzahl aufgerufener Dokumente	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	M05 / M01
V02	Anzahl aufgerufener relevanter Dokumente / Anzahl aufgerufener Dokumente	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	M06 / M01
V03	Anzahl aufgerufener relevanter Dokumente / Anzahl relevanter Dokumente im Korpus	Anz. aufg. rel. Dok. / Anz. rel. Dok. im Korpus	M06 / REL01
V04	Anzahl aufgerufener relevanter Dokumente / Anzahl zurückgegebener relevanter Dokumente	Anz. aufg. rel. Dok. / Anz. zurückgeg. rel. Dok.	M06 / REL02
V05	Anzahl falsch irrelevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	M07 / M01
V06	Anzahl falsch irrelevant bewerteter Dokumente / Anzahl irrelevant bewerteter Dokumente	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	M07 / M09
V07	Anzahl falsch irrelevant bewerteter Dokumente / Anzahl richtig irrelevant bewerteter Dokumente	Anz. falsch irrel. bew. Dok. / Anz. richtig irrel. bew. Dok.	M07 / M15
V08	Anzahl falsch relevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. falsch rel. bew. Dok. / Anz. aufg. Dok.	M08 / M01
V09	Anzahl falsch relevant bewerteter Dokumente / Anzahl relevant bewerteter Dokumente	Anz. falsch rel. bew. Dok. / Anz. rel. bew. Dok.	M08 / M10
V10	Anzahl falsch relevant bewerteter Dokumente / Anzahl richtig relevant bewerteter Dokumente	Anz. falsch rel. bew. Dok. / Anz. richtig rel. bew. Dok.	M08 / M16
V11	Anzahl irrelevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. irrel. bew. Dok. / Anz. aufg. Dok.	M09 / M01
V12	Anzahl relevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. rel. bew. Dok. / Anz. aufg. Dok.	M10 / M01
V13	Anzahl richtig bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. richtig bew. Dok. / Anz. aufg. Dok.	M14 / M01
V14	Anzahl richtig irrelevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	M15 / M01
V15	Anzahl richtig irrelevant bewerteter Dokumente / Anzahl aufgerufener irrelevanter Dokumente	Anz. richtig irrel. bew. Dok. / Anz. aufg. irrel. Dok.	M15 / M05

Fortsetzung auf nächster Seite

Tab. B.12 (Fortsetzung) Variablengrp. 4

ID	Beschreibung	Kurzform	Formel
V16	Anzahl richtig irrelevant bewerteter Dokumente / Anzahl falsch irrelevant bewerteter Dokumente	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.	M15 / M07
V17	Anzahl richtig irrelevant bewerteter Dokumente / Anzahl irrelevant bewerteter Dokumente	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	M15 / M09
V18	Anzahl richtig relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente / Anzahl aufgerufener Dokumente der ersten zehn angezeigten Dokumente	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. aufg. Dok. (erste 10 Dok.)	M17 / M02
V19	Anzahl richtig relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente / Anzahl relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)	M17 / M11
V20	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche / Anzahl aufgerufener Dokumente der ersten durchgeführten Suche	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. aufg. Dok. (erste Suche)	M18 / M03
V21	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche / Anzahl relevant bewerteter Dokumente der ersten durchgeführten Suche	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. bew. Dok. (erste Suche)	M18 / M12
V22	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche / Anzahl relevanter Dokumente im Korpus	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. Dok. im Korpus	M18 / REL01
V23	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche / Anzahl zurückgegebener relevanter Dokumente der ersten durchgeführten Suche	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. zurückgeg. rel. Dok. (erste Suche)	M18 / REL03
V24	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche / Anzahl aufgerufener Dokumente der letzten durchgeführten Suche	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. aufg. Dok. (letzte Suche)	M19 / M04
V25	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche / Anzahl relevant bewerteter Dokumente der letzten durchgeführten Suche	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. bew. Dok. (letzte Suche)	M19 / M13
V26	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche / Anzahl relevanter Dokumente im Korpus	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. Dok. im Korpus	M19 / REL01
V27	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche / Anzahl zurückgegebener relevanter Dokumente der letzten durchgeführten Suche	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. zurückgeg. rel. Dok. (letzte Suche)	M19 / REL04
V28/PCP	Anzahl richtig relevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.	M16 / M01
V29	Anzahl richtig relevant bewerteter Dokumente / Anzahl aufgerufener relevanter Dokumente	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	M16 / M06
V30	Anzahl richtig relevant bewerteter Dokumente / Anzahl falsch relevant bewerteter Dokumente	Anz. richtig rel. bew. Dok. / Anz. falsch rel. bew. Dok.	M16 / M08
V31/BP	Anzahl richtig relevant bewerteter Dokumente / Anzahl relevant bewerteter Dokumente	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.	M16 / M10
V32/BR	Anzahl richtig relevant bewerteter Dokumente / Anzahl relevanter Dokumente im Korpus	Anz. richtig rel. bew. Dok. / Anz. rel. Dok. im Korpus	M16 / REL01
V33	Anzahl richtig relevant bewerteter Dokumente / Anzahl zurückgegebener relevanter Dokumente	Anz. richtig rel. bew. Dok. / Anz. zurückgeg. rel. Dok.	M16 / REL02

Tab. B.13.: Sonstige Leistungsmaße bei binärer Relevanzskala (Variablengrp. 5). Für S05 wird jeweils der erste Aufruf eines Dokuments berücksichtigt.

ID	Beschreibung	Kurzform
S01	Anzahl Suchen	Anz. Suchen
S02	Erste betrachtete Rankingposition	Erste betr. Rankingpos.
S03	Letzte betrachtete Rankingposition	Letzte betr. Rankingpos.
S04	Suchdauer	Suchdauer
S05/S05-log[a]	Zeit bis zum ersten richtig relevant bewerteten Dokument	Zeit zum ersten richtig rel. bew. Dok.

[a] S05 entspricht dem im ersten Experiment verwendeten Benutzerleistungsmaß T1.

B.10. Kovariaten zur statistischen Kontrolle personenbezogener Störfaktoren

Im Folgenden wird eine Übersicht über die im zweiten Experiment berücksichtigten Kovariaten gegeben. Diese lassen sich in drei Gruppen untergliedern: demographische Kovariaten, erfahrungsbezogene Kovariaten sowie eine motivationsbezogene Kovariate. Mit Ausnahme von den Ergebnissen der Wissenstests (K06 u. K11) handelt es sich bei den betrachteten Kovariaten ausschließlich um Selbstauskunftsmaße.

Tab. B.14.: Übersicht über berücksichtigte Kovariaten.

ID	Gruppe	Beschreibung
K01	Person	Geburtsjahr
K02	Person	Geschlecht
K03	Person	Muttersprache
K04	Erfahrung	Computernutzungsjahre
K05	Erfahrung	Computernutzungsstunden
K06	Erfahrung	Domänenwissen
K07	Erfahrung	Selbsteinschätzung Domänenwissen
K08	Erfahrung	Selbsteinschätzung Suchmaschinenwissen
K09	Erfahrung	Suchmaschinnutzungsjahre
K10	Erfahrung	Suchmaschinennutzungsstunden
K11	Erfahrung	Suchmaschinenwissen
K12	Motivation	Ausgangsmotivation

C. Weitere Ergebnisse zu Experiment 2

Dieser Anhang enthält weitere Ergebnisse, die im Rahmen der Auswertung des zweiten Experiments entstanden sind. Auf eine Darstellung innerhalb der Arbeit wird aus Gründen der Übersichtlichkeit verzichtet, es wird jedoch in vielen Fällen darauf verwiesen. Die Struktur des Kapitels ist im Wesentlichen an die Struktur von Kapitel 6 angelehnt, wobei die vertiefenderen Ergebnisse der Item- und Faktorenanalyse den weitergehenden Ergebnissen der durchgeführten Varianzanalysen vorangestellt sind. Die letzten beiden Abschnitte fassen weitere Ergebnisse in Bezug auf die Überprüfung der Gütekriterien des zweiten Experiments zusammen.

C.1. Weitere Ergebnisse der Itemanalyse

Neben den in Abschnitt 6.4.4.1 berichteten Ergebnissen der Trennschärfeanalyse über alle in Experiment 2 verwendeten Zufriedenheitsitems, wird die Trennschärfe im Folgenden zusätzlich separat für die EUCS- und Zusatzitems bestimmt. Während die Trennschärfe aller Zusatzitems als zufriedenstellend bezeichnet werden kann (vgl. Tab. C.2), erweisen sich zwei der EUCS-Items (F09 u. F11) im Zuge der Itemanalyse als weniger trennscharf (vgl. Tab. C.1).

Tab. C.1.: Trennschärfe der EUCS-Items ($n = 240$).

Item	Beschreibung	Korrigierte Item-Total-Korrelation
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	0,77
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	0,74
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	0,70
F04	Liefert die Suchmaschine genügend Information?	0,73
F05	Ist die Suchmaschine präzise?	0,79
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	0,84
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,58
F08	Ist die Suchmaschine benutzerfreundlich?	0,61
F09	Ist die Suchmaschine einfach zu bedienen?	**0,44**
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	0,72
F11	Liefert die Suchmaschine aktuelle Information?	**0,49**

Tab. C.2.: Trennschärfe der Zusatzsitems ($n = 240$).

Item	Beschreibung	Korrigierte Item-Total-Korrelation
F14	Es war einfach, die Aufgabe zu bearbeiten.	0,66
F15	Es war einfach, zu dem Thema zu suchen.	0,64
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	0,79
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	0,60
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	0,74
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	0,77
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,69
F21	Ich hatte genügend Zeit, um eine effektive Suche durchzuführen.	0,53
F23	Ich bin mit meiner Suchleistung zufrieden.	0,71
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	0,60

C.2. Weitere Ergebnisse der Faktorenanalyse

In diesem Abschnitt werden weitergehende Ergebnisse bezüglich der explorativen Faktorenanalysen beschrieben. Während innerhalb der Arbeit lediglich die Endresultate der einzelnen Faktorenanalysen berichtet werden, findet sich im Folgenden ein vertiefender Einblick in die Herleitung der einzelnen Skalen. Analog zu Abschnitt 6.4.4.2 befasst sich der erste Unterabschnitt zunächst mit dem Replikationsversuch der fünf im EUCS-Instrument enthaltenen Zufriedenheitsskalen. Der zweite Unterabschnitt widmet sich der Analyse der zusätzlich entwickelten Zufriedenheitsitems. Der dritte Unterabschnitt schließlich beschreibt die gemeinsame Analyse von EUCS- und Zusatzitems. Im letzten Unterabschnitt sind ergänzend zu den in Abschnitt 6.4.4.3 berichteten Resultaten der Reliabilitäts- und Validitätsanalyse weitere Ergebnisse in Bezug auf die in Experiment 2 betrachteten kritischen Fallgruppen angegeben.

C.2.1. Analyse der EUCS-Items

Zur Analyse der EUCS-Items werden sieben Hauptkomponentenanalysen durchgeführt, deren Ergebnisse im Folgenden dargestellt sind. Da es sich um einen Replikationsversuch handelt, werden für die Durchführung der Analyse dieselbe Faktorenanzahl und dasselbe Rotationsverfahren wie bei Doll und Torkzadeh (1988) angewendet: Hauptkomponentenanalyse mit 5 Faktoren und anschließender orthogonaler Varimax-Rotation (vgl. Tab. C.3). Die resultierende Faktorlösung ist allerdings noch nicht zufriedenstellend, da Item F07 ursprünglich der Skala *Darstellung* zugeordnet wird, deren zweites Item jedoch im Rahmen der vorliegenden Untersuchung nicht mit erhoben wird (vgl. Abschn. 6.3.2). Deshalb wird die Analyse im nächsten Schritt um einen Faktor reduziert (vgl. Tab. C.4). Anschließend wird Item F09 von der Analyse ausgeschlossen (vgl. Tab. C.5), weil es in beiden Analysen allein auf einen Faktor lädt und im Rahmen der Itemanalyse bereits mehrfach aufgrund seiner unzureichenden Trennschärfe auffällt (vgl. Tab. C.1 sowie 6.11).

Allerdings führen auch diese beiden zusätzlich durchgeführten Analysen zu keiner gut interpretierbaren Faktorlösung. Die Tatsache, dass in allen drei Lösungen Doppelladungen einzelner Items auf zwei oder mehr Faktoren auftreten legt den Schluss nahe, dass eine oblique Faktorrotation, bei der Korrelationen zwischen den Faktoren erlaubt sind, den Daten besser gerecht wird. Deshalb wird im nächsten Schritt eine Hauptkomponentenanalyse mit obliquer Rotation (oblimin) und 4 Faktoren durchgeführt (vgl. Tab. C.6).

In einer weiteren Analyse werden die Items F10 und F11 ausgeschlossen. Die Tatsache, dass Item F10 im Rahmen der bisherigen Analysen häufig gemeinsam mit Items auf einen Faktor lädt, die die Zufriedenheit mit Inhalt und Genauigkeit adressieren, führt zu der Annahme, dass das Wort *rechtzeitig* von einigen Probanden überlesen wurde und F10 somit eine völlig neue Bedeutung erhält. Da eine Skala weiterhin mindestens zwei Items enthalten sollte und F11 nunmehr das einzige Item ist, das noch die Aktualität der Suchergebnisse beinhaltet, wird auch

Tab. C.3.: PCA 1: EUCS-Items mit 5 Faktoren und Varimax-Rotation (alle 11 EUCS-Items). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,84**	0,06	0,21	0,11	0,15
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,82**	0,22	0,21	0,04	0,13
F04	Liefert die Suchmaschine genügend Information?	**0,79**	0,25	0,12	0,09	0,22
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,72**	0,16	**0,52**	0,05	0,00
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,65**	**0,30**	**0,50**	0,17	0,11
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,22	**0,87**	0,21	0,03	0,10
F08	Ist die Suchmaschine benutzerfreundlich?	0,22	**0,77**	0,16	**0,36**	0,14
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	**0,30**	0,27	**0,75**	0,12	**0,31**
F05	Ist die Suchmaschine präzise?	**0,57**	0,21	**0,65**	0,21	0,02
F09	Ist die Suchmaschine einfach zu bedienen?	0,12	0,20	0,14	**0,93**	0,15
F11	Liefert die Suchmaschine aktuelle Information?	0,22	0,16	0,16	0,16	**0,92**

Tab. C.4.: PCA 2: EUCS-Items mit 4 Faktoren und Varimax-Rotation (alle 11 EUCS-Items). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,88**	0,20	0,08	0,024
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,84**	0,04	0,08	0,18
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,82**	0,20	0,01	0,17
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,80**	**0,33**	0,20	0,13
F05	Ist die Suchmaschine präzise?	**0,79**	0,27	0,27	0,03
F04	Liefert die Suchmaschine genügend Information?	**0,75**	0,22	0,04	0,26
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	**0,58**	**0,37**	0,21	**0,31**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,27	**0,88**	0,03	0,11
F08	Ist die Suchmaschine benutzerfreundlich?	0,25	**0,77**	**0,35**	0,15
F09	Ist die Suchmaschine einfach zu bedienen?	0,15	0,20	**0,93**	0,16
F11	Liefert die Suchmaschine aktuelle Information?	0,24	0,17	0,16	**0,92**

diese Frage von der weiteren Analyse ausgeschlossen und die Faktorenanzahl erneut um einen Faktor reduziert (vgl. Tab. C.7).

Da die drei Items F07, F08 und F09 inhaltlich sehr nahe beieinander liegen, wird die Analyse im nächsten Schritt um einen weiteren Faktor auf 2 Faktoren reduziert (vgl. Tab. C.8). Diese Maßnahme stellt zunächst den Schlusspunkt der Analyse der EUCS-Items dar. Die resultierende Zweifaktorenlösung erklärt 71,76 % der Gesamtvarianz. Eine erste Reliabilitätsbeurteilung der Skalen ergibt jedoch, dass sich der Reliabilitätskoeffizient Cronbachs Alpha von 0,74 auf 0,77 erhöhen würde, wenn man Item F09 ebenfalls entfernen würde. Durch den Ausschluss von Item F09 wird schließlich auch ein Aufspalten des ersten Faktors in die durch das EUCS-Instrument vorgeschlagenen Teilskalen Inhalt und Genauigkeit möglich. Die resultierende Dreifaktorenlösung ist in Abschnitt 6.4.4.2 detailliert erläutert.

C.2.2. Analyse der Zusatzitems

Zur Analyse der Zusatzitems wird zunächst eine Hauptkomponentenanalyse mit anschließender Oblimin-Rotation eingesetzt. Der Scree-Plot für diese Analyse ist in Abbildung C.1 dargestellt. Der Knick an der vierten Hauptkomponente ist nach dem Ellenbogenkriterium ein Indiz für eine Dreifaktorenlösung.

Tabelle C.9 zeigt die rotierte Ladungsmatrix dieser Lösung. Auch diese Faktorenlösung ist aus inhaltlicher Sicht zufriedenstellend. Unklar sind jedoch die eng beieinander liegende Doppelladungen von Item F16 und F18, die letztlich zu der Entscheidung für die in Abschnitt 6.4.4.2 berichteten Vierfaktorenlösung führen. Diese letztlich ergriffene Lösung ist auch aufgrund des

Tab. C.5.: PCA 3: EUCS-Items mit 4 Faktoren und Varimax-Rotation (EUCS-Items ohne F09). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,82**	0,28	0,08	0,16
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,81**	0,26	0,22	0,12
F04	Liefert die Suchmaschine genügend Information?	**0,78**	0,18	0,26	0,22
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,68**	**0,58**	0,17	0,01
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,60**	**0,57**	**0,33**	0,13
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	0,25	**0,75**	0,29	**0,32**
F05	Ist die Suchmaschine präzise?	**0,51**	**0,71**	0,25	0,05
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,21	0,19	**0,86**	0,08
F08	Ist die Suchmaschine benutzerfreundlich?	0,18	0,22	**0,83**	0,18
F11	Liefert die Suchmaschine aktuelle Information?	0,21	0,16	0,18	**0,93**

Tab. C.6.: PCA 4: EUCS-Items mit 4 Faktoren und Oblimin-Rotation (alle 11 EUCS-Items). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,93**	0,01	−0,01	−0,10
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,90**	−0,17	−0,01	0,09
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,84**	0,03	−0,10	0,07
F05	Ist die Suchmaschine präzise?	**0,80**	0,08	0,18	−0,11
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,78**	0,15	0,10	0,00
F04	Liefert die Suchmaschine genügend Information?	**0,74**	0,05	−0,07	0,17
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	**0,49**	0,22	0,10	0,22
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,04	**0,94**	−0,11	0,02
F08	Ist die Suchmaschine benutzerfreundlich?	0,01	**0,75**	0,25	0,04
F09	Ist die Suchmaschine einfach zu bedienen?	0,02	0,00	**0,96**	0,05
F11	Liefert die Suchmaschine aktuelle Information?	0,01	0,02	0,04	**0,96**

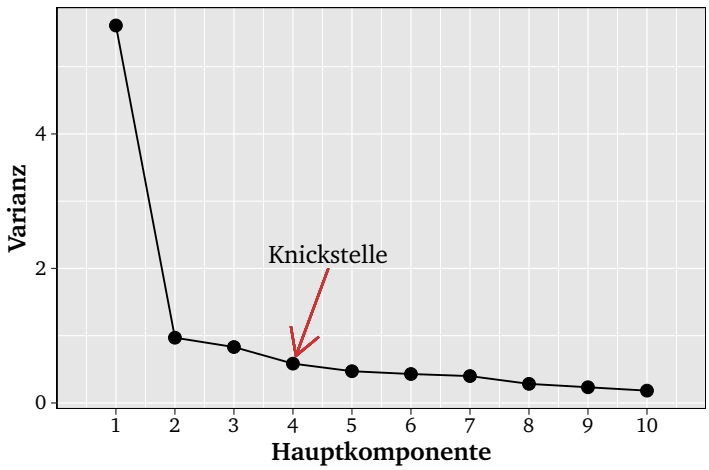


Abb. C.1.: Scree-Plot für Hauptkomponenetenanalyse der Zusatzitems.

erklärten Varianzanteils zu bevorzugen.

C.2.3. Analyse aller Items

Insgesamt werden zur Analyse aller Zufriedenheitsitems drei Hauptkomponentenanalysen mit drei, vier und fünf Faktoren gerechnet. Die Tabellen C.10 und C.11 zeigen die Ladungsmuster der beiden als weniger geeignet eingestuften Lösungen. Die Gründe sich gegen diese beiden Faktorlösungen zu entscheiden, werden im Folgenden kurz erläutert. Wie im Vergleich der Tabellen C.10 und C.11 zu erkennen, unterscheiden sich die Drei- und Vierfaktorenlösung im Wesentlichen

Tab. C.7.: PCA 5: EUCS-Items mit 3 Faktoren und Oblimin-Rotation (EUCS-Items ohne F10 u. F11). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,93**	−0,16	−0,01
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,90**	0,00	−0,04
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,87**	0,03	−0,08
F04	Liefert die Suchmaschine genügend Information?	**0,79**	0,07	−0,02
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,78**	0,15	0,09
F05	Ist die Suchmaschine präzise?	**0,77**	0,07	0,14
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,04	**0,95**	−0,10
F08	Ist die Suchmaschine benutzerfreundlich?	0,02	**0,75**	0,27
F09	Ist die Suchmaschine einfach zu bedienen?	0,02	0,01	**0,98**

Tab. C.8.: PCA 6: EUCS-Items mit 2 Faktoren und Oblimin-Rotation (EUCS-Items ohne F10 u. F11). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,91**	−0,12
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,91**	−0,04
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,88**	−0,04
F04	Liefert die Suchmaschine genügend Information?	**0,80**	0,04
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,78**	0,20
F05	Ist die Suchmaschine präzise?	**0,76**	0,17
F08	Ist die Suchmaschine benutzerfreundlich?	0,05	**0,85**
F09	Ist die Suchmaschine einfach zu bedienen?	−0,10	**0,81**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,14	**0,71**

dadurch, dass die Zufriedenheit mit der eigenen Leistung und der gestellten Aufgabe zusammengefasst werden. Gegen die Dreifaktorenlösung spricht aus diesem Grund vor allem der Wunsch nach einer weiteren Differenzierung in Bezug auf diese beiden Dimensionen.

Die in Tabelle C.11 dargestellte Fünffaktorenlösung unterscheidet sich insbesondere durch die Entstehung einer zweiten die Benutzerfreundlichkeit der Suchmaschine beurteilenden Komponente. Da zudem die aufgrund der zuvor durchgeführten Einzelanalysen erwartete Aufspaltung der die Qualität der Sucherfahrung beschreibenden Komponente in die aus dem EUCS-Instrument replizierten Teilskalen Inhalt und Genauigkeit nicht auftritt, wird sich schließlich für die Vierfaktorenlösung entschieden.

C.2.4. Reliabilitäts- und Validitätsanalyse

Die Tabellen C.12 und C.13 zeigen die Reliabilitäts- und Validitätskoeffizienten der im zweiten Experiment ermittelten Skalen für die Gesamtstichprobe, nachdem zunächst jeweils eine der kritischen Fallgruppen SP_{SB}, SP_{MV} und SP_{TD} von der Stichprobe SP_A ausgeschlossen wurde. Wie im Fall der in Abschnitt 6.4.4.3 berichteten Befunde, ergeben sich auch zwischen den hier betrachteten Fallgruppen nur geringe Differenzen zwischen den ermittelten Koeffizienten, sodass davon auszugehen ist, dass die Hinzunahme der hier ausgeschlossenen Fälle kaum Auswirkungen auf die Ergebnisse haben wird.

In den Tabellen C.14 und C.15 sind die Reliabilitäts- und Validitätskoeffizienten der Originalskalen des EUCS-Instruments aufgeführt. Es zeigt sich, dass die Gütekriterien mit Ausnahme der Skalen *Ease of use* und *Timeliness* erfüllt sind. So weisen die Skalen *Content* und *Accuracy* für beide Aufgaben interne Konsistenzen größer $0{,}7$ auf. Zudem wird der kritische Schwellenwert der Kriteriumsvalidität ($> 0{,}4$) für beide Aufgaben bei allen Skalen überschritten. Die niedrigeren Reliabilitätskoeffizienten der Skalen *Ease of use* und *Timeliness* lassen sich möglicherweise durch den Umstand erklären, dass diese Skalen im Rahmen des vorliegenden Experiments nicht vollständig repliziert werden konnten. Außerdem lässt die Tatsache, dass die Werte für die zweite Aufgabe besser ausfallen vermuten, dass die zu diesen Skalen beitragen Frageitems erst nach einer gewissen Einarbeitungsphase zuverlässig beantwortet werden können. Es ist jedoch auch

Tab. C.9.: PCA 1: Zusatzitems mit 3 Faktoren und Oblimin-Rotation (alle 10 Zusatzitems). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	**0,90**	−0,04	−0,10
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	**0,75**	−0,22	0,26
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,74**	0,16	0,09
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	**0,62**	**0,37**	0,04
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,57**	**0,40**	0,00
F14	Es war einfach, die Aufgabe zu bearbeiten.	−0,05	**0,83**	0,17
F15	Es war einfach, zu dem Thema zu suchen.	0,11	**0,83**	−0,03
F21	Ich hatte genügend Zeit, um eine effektive Suche durchzuführen.	−0,04	−0,06	**0,94**
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,18	0,16	**0,64**
F23	Ich bin mit meiner Suchleistung zufrieden.	0,10	0,29	**0,62**

Tab. C.10.: PCA 1: Alle Items mit 3 Faktoren und Oblimin-Rotation (alle 16 Items). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,95**	−0,06	−0,13
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,90**	−0,01	−0,03
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,82**	0,05	−0,01
F05	Ist die Suchmaschine präzise?	**0,78**	−0,03	0,13
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,75**	0,17	−0,03
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,71**	0,05	0,17
F04	Liefert die Suchmaschine genügend Information?	**0,70**	0,10	0,04
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	**0,57**	0,24	0,20
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	**0,54**	−0,07	**0,34**
F14	Es war einfach, die Aufgabe zu bearbeiten.	−0,07	**0,93**	0,00
F15	Es war einfach, zu dem Thema zu suchen.	0,14	**0,73**	−0,04
F23	Ich bin mit meiner Suchleistung zufrieden.	0,09	**0,72**	0,03
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,14	**0,59**	0,12
F08	Ist die Suchmaschine benutzerfreundlich?	0,01	−0,03	**0,88**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,04	−0,02	**0,82**
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	0,00	0,19	**0,70**

im Fall der Originalskalen festzustellen, dass die Werte für die betrachteten Fallgruppen erneut sehr nah beieinander liegen.

Tab. C.11.: PCA 2: Alle Items mit 5 Faktoren und Oblimin-Rotation (alle 16 Items). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,87**	−0,06	−0,05	0,08	0,01
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,81**	0,04	0,14	−0,09	0,06
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,74**	0,03	−0,02	0,18	0,02
F04	Liefert die Suchmaschine genügend Information?	**0,71**	0,21	0,09	0,08	−0,21
F05	Ist die Suchmaschine präzise?	**0,69**	0,15	0,02	0,03	0,10
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,63**	−0,06	0,17	0,11	0,18
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,53**	0,03	0,04	0,13	**0,38**
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	**0,45**	0,15	0,26	0,08	0,20
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,09	**0,87**	−0,04	0,06	−0,08
F08	Ist die Suchmaschine benutzerfreundlich?	0,00	**0,85**	0,09	−0,09	0,10
F15	Es war einfach, zu dem Thema zu suchen.	0,11	0,00	**0,89**	−0,09	0,01
F14	Es war einfach, die Aufgabe zu bearbeiten.	−0,08	0,02	**0,87**	0,17	−0,03
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,14	0,08	0,00	**0,80**	−0,06
F23	Ich bin mit meiner Suchleistung zufrieden.	0,05	−0,08	0,15	**0,79**	0,07
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	0,25	0,03	0,06	−0,03	**0,73**
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	−0,16	**0,41**	0,03	**0,30**	**0,52**

Tab. C.12.: Skalenreliabilität und Kriteriumsvalidität unter Ausschluss kritischer Fallgruppen für $A1$. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	Benutzerfreundlichkeit	Cronbachs Alpha			Kriteriumsvalidität		
		$SP_A \setminus SP_{SB}$ $n = 64$	$SP_A \setminus SP_{MV}$ $n = 115$	$SP_A \setminus SP_{TD}$ $n = 103$	$SP_A \setminus SP_{SB}$ $n = 64$	$SP_A \setminus SP_{MV}$ $n = 115$	$SP_A \setminus SP_{TD}$ $n = 103$
SK01	Inhalt	0,9	0,9	0,9	0,82	0,84	0,84
SK02	Genauigkeit	0,86	0,85	0,83	0,76	0,77	0,76
SK03	Benutzerfreundlichkeit	0,81	0,82	0,79	0,71	0,66	0,64
SK04	Suche	0,89	0,86	0,86	0,83	0,85	0,85
SK05/10	Aufgabe	0,76	0,77	0,77	0,61	0,57	0,59
SK06	Eigenleistung	0,85	0,82	0,81	0,49	0,51	0,52
SK07	Benutzerfreundlichkeit	0,71	0,68	0,69	0,65	0,65	0,66
SK08	Suche	0,93	0,93	0,92	0,86	0,88	0,87
SK09	Benutzerfreundlichkeit	0,82	0,82	0,81	0,75	0,7	0,7
SK11	Eigenleistung	0,82	0,77	0,77	0,54	0,56	0,58

Tab. C.13.: Skalenreliabilität und Kriteriumsvalidität unter Ausschluss kritischer Fallgruppen für $A2$. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	Beschreibung	Cronbachs Alpha			Kriteriumsvalidität		
		$SP_A \setminus SP_{SB}$ $n = 66$	$SP_A \setminus SP_{MV}$ $n = 117$	$SP_A \setminus SP_{TD}$ $n = 104$	$SP_A \setminus SP_{SB}$ $n = 66$	$SP_A \setminus SP_{MV}$ $n = 117$	$SP_A \setminus SP_{TD}$ $n = 104$
SK01	Inhalt	0,91	0,92	0,91	0,85	0,84	0,82
SK02	Genauigkeit	0,85	0,87	0,86	0,81	0,86	0,85
SK03	Benutzerfreundlichkeit	0,68	0,69	0,67	0,64	0,64	0,59
SK04	Suche	0,94	0,93	0,92	0,82	0,81	0,8
SK05/10	Aufgabe	0,83	0,86	0,87	0,5	0,58	0,56
SK06	Eigenleistung	0,79	0,76	0,73	0,48	0,55	0,51
SK07	Benutzerfreundlichkeit	0,62	0,66	0,64	0,75	0,71	0,71
SK08	Suche	0,95	0,95	0,95	0,87	0,88	0,87
SK09	Benutzerfreundlichkeit	0,78	0,74	0,73	0,7	0,69	0,65
SK11	Eigenleistung	0,78	0,75	0,74	0,52	0,62	0,57

Tab. C.14.: Skalenreliabilität der Originalskalen des EUCS-Instruments nach Datenqualität und unter Ausschluss kritischer Fallgruppen. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	**Cronbachs Alpha**				
	SP_A	**SP_B**	**$SP_A \setminus SP_{SB}$**	**$SP_A \setminus SP_{MV}$**	**$SP_A \setminus SP_{TD}$**
			Aufgabe 1		
	$n = 119$	$n = 54$	$n = 64$	$n = 115$	$n = 103$
Content	0,89	0,88	0,89	0,89	0,88
Accuracy	0,89	0,89	0,89	0,89	0,88
Ease of use	0,6	0,59	0,6	0,62	0,6
Timeliness	0,56	0,64	0,64	0,56	0,56
			Aufgabe 2		
	$n = 121$	$n = 55$	$n = 66$	$n = 117$	$n = 104$
Content	0,9	0,88	0,89	0,9	0,9
Accuracy	0,89	0,89	0,9	0,9	0,87
Ease of use	0,73	0,68	0,72	0,74	0,67
Timeliness	0,69	0,71	0,63	0,69	0,73

Tab. C.15.: Kriteriumsvalidität der Originalskalen des EUCS-Instruments nach Datenqualität und unter Ausschluss kritischer Fallgruppen. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	**Kriteriumsvalidität**				
	SP_A	**SP_B**	**$SP_A \setminus SP_{SB}$**	**$SP_A \setminus SP_{MV}$**	**$SP_A \setminus SP_{TD}$**
			Aufgabe 1		
	$n = 119$	$n = 54$	$n = 64$	$n = 115$	$n = 103$
Content	0,81	0,82	0,8	0,81	0,81
Accuracy	0,8	0,76	0,78	0,81	0,79
Ease of use	0,59	0,63	0,65	0,59	0,58
Timeliness	0,69	0,81	0,81	0,7	0,7
			Aufgabe 2		
	$n = 121$	$n = 55$	$n = 66$	$n = 117$	$n = 104$
Content	0,87	0,82	0,85	0,87	0,87
Accuracy	0,82	0,8	0,81	0,82	0,8
Ease of use	0,6	0,46	0,56	0,6	0,53
Timeliness	0,7	0,66	0,69	0,69	0,66

C.3. Weitere Ergebnisse der Varianzanalysen

In diesem Anhang sind weitere Ergebnisse zu den im Rahmen des zweiten Experiments durchgeführten Varianzanalysen zusammengestellt auf deren Darstellung innerhalb des Hauptteils der Arbeit aus Gründen der Übersichtlichkeit verzichtet wird. Der vorliegende Abschnitt ist in drei Teile gegliedert. Abschnitt C.3.1 enthält zunächst eine Übersicht über diejenigen Variablen die in keiner der fünf Stichproben einen signifikanten Effekt der untersuchten Einflussgrößen zeigen. Abschnitt C.3.2 umfasst die Mittelwerte der Interaktionen für die in den Abschnitten 6.4.3 und 6.4.5 berichteten signifikanten Ergebnisse der Varianzanalysen. Im Anschluss daran geben die in Abschnitt C.3.3 dargestellten Tabellen Aufschluss über das im Einzelnen verwendete Analyseverfahren, die Stabilität der berichteten Effekte (eindeutig vs. tendenziell) sowie das Signifikanzniveau der jeweiligen unabhängigen Variablen.

C.3.1. Variablen ohne signifikante Unterschiede

Tabelle C.16 stellt eine Übersicht derjenigen Variablen bereit, die in keiner der fünf Stichproben einen signifikanten Effekt der untersuchten Einflussgrößen zeigen. Das Fehlen eines Effekts wird aufgeschlüsselt nach den beiden Aufgaben und den beiden Datenqualitätsstufen SP_A und SP_B. Darüber hinaus gibt die Tabelle die Stichprobengröße, den Grad der Kontrolle des Datensatzes (Topicbalancierung ja/nein) sowie die Art der Varianzanalyse (klassisch vs. robust) an. Minuszeichen ($-$) bedeuten, dass für diese Teilstichprobe keine ausreichende Fallanzahl für eine Analyse zur Verfügung steht, Pluszeichen hingegen ($+$), dass signifikante Effekte der unabhängigen Variable vorliegen. Des Weiteren kennzeichnen Sterne ($*$) Variablen, für die keine eindeutige Aussage getroffen werden kann, da zwischen einer und drei Zufallsstichproben signifikante Abhängigkeiten aufweisen. Im Fall der Erwartungsitems weisen Kreise ($\circ$) darauf hin, dass diese Variablen ausschließlich im Anschluss an die zweite Suchaufgabe erhoben werden.

Tab. C.16.: **neu** Übersicht über Variablen ohne signifikante Unterschiede.

			Aufgabe 1				**Aufgabe 2**	
			SP_A		**SP_B**		**SP_A**	
ID	**Beschreibung**	**topic**	**n**	**V**	**n**	**V**	**n**	**V**
M01	Anz. aufg. Dok.	nein	108	R	*	*	96	R
M02	Anz. aufg. Dok. (erste 10 Dok.)	ja	96	K[b]/R	–	–	*	*
M03	Anz. aufg. Dok. (erste Suche)	nein	108	R	40	K/R	+	+
M04	Anz. aufg. Dok. (letzte Suche)	nein	108	R	40	K/R	*	*
M06	Anz. aufg. rel. Dok.	nein	108	R	40	K/R	*	*
M09	Anz. irrel. bew. Dok.	ja	96	R	–	–	*	*
M10	Anz. rel. bew. Dok.	ja	96	R	–	–	80	R
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	ja	96	K/R	–	–	+	+
M12	Anz. rel. bew. Dok. (erste Suche)	ja	96	R	–	–	+	+
M13	Anz. rel. bew. Dok. (letzte Suche)	ja	96	R	–	–	*	*
M15	Anz. richtig irrel. bew. Dok.	nein	+	+	40	R	+	+
M16	Anz. richtig rel. bew. Dok.	ja	96	R	–	–	*	*
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	nein	108	R	40	K/R	+	+
M18	Anz. richtig rel. bew. Dok. (erste Suche)	nein	108	R	*	*	+	+
M19	Anz. richtig rel. bew. Dok. (letzte Suche)	ja	96	R	–	–	*	*
B04	Durchschn. Bew. rel. Dok.	ja	*	*	–	–	80	K/R[a]
B05	Durchschn. Bew. rel. Dok. (erste Suche)	ja	88	R	–	–	–	-
Z01	Durchschn. Betrachtungsz. aller Dok.	nein	108	R	40	R	96	K[b]/R
Z01-log	Durchschn. Betrachtungsz. aller Dok.	ja	*	*	–	–	80	K
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	nein	96	R	40	R	96	R
Z02-log	Durchschn. Betrachtungsz. falsch bew. Dok.	ja	96	R	–	–	*	*

[a] Nur einmal robuste Analyse.

[b] Nur einmal klassische Analyse.

Fortsetzung auf nächster Seite

Tab. C.16 (Fortsetzung)

			Aufgabe 1				Aufgabe 2	
			SP_A		SP_B		SP_A	
ID	**Beschreibung**	**topic**	**n**	**V**	**n**	**V**	**n**	**V**
Z03-log	Durchschn. Betrachtungsz. falsch irrel. bew. Dok.	ja	–	–	–	–	80	K
Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	ja	96	R	–	–	80	R
Z05-log	Durchschn. Betrachtungsz. irrel. bew. Dok.	ja	96	K	–	–	80	K/R
Z06	Durchschn. Betrachtungsz. irrel. Dok.	nein	92	R	–	–	–	-
Z06-log	Durchschn. Betrachtungsz. irrel. Dok.	nein	92	R	–	–	–	-
Z07	Durchschn. Betrachtungsz. rel. bew. Dok.	ja	96	R	–	–	80	R
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	ja	96	K	–	–	80	K/R[b]
Z08	Durchschn. Betrachtungsz. rel. Dok.	ja	96	R	–	–	80	R
Z08-log	Durchschn. Betrachtungsz. rel. Dok.	ja	96	K	–	–	80	K/R
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	ja	96	R	–	–	*	*
Z09-log	Durchschn. Betrachtungsz. richtig bew. Dok.	nein	108	K	40	K	96	R
Z10	Durchschn. Betrachtungsz. richtig irrel. bew. Dok.	nein	80	R	–	–	–	-
Z10-log	Durchschn. Betrachtungsz. richtig irrel. bew. Dok.	nein	80	K/R	–	–	–	-
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	ja	96	R	–	–	80	R
Z11-log	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	ja	96	K	–	–	80	K/R[b]
V03	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	nein	108	R	40	K/R[a]	*	*
V04	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	nein	108	R	40	K	*	*
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	ja	96	K	–	–	*	*
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	ja	96	K/R[a]	–	–	*	*
V22	$\frac{\text{Anz. richtig rel. bew. Dok. (erste Suche)}}{\text{Anz. rel. Dok. im Korpus}}$	ja	88	R	–	–	–	-
V23	$\frac{\text{Anz. richtig rel. bew. Dok. (erste Suche)}}{\text{Anz. zurückgeg. rel. Dok. (erste Suche)}}$	ja	88	R	–	–	–	-
V24	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. aufg. Dok. (letzte Suche)}}$	ja	88	K	–	–	–	-
V26	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. rel. Dok. im Korpus}}$	ja	88	R	–	–	–	-
V27	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. zurückgeg. rel. Dok. (letzte Suche)}}$	ja	88	R	–	–	–	-
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	ja	*	*	–	–	80	K/R[a]
V32/BR	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	ja	96	R	–	–	*	*
V33	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	ja	96	R	–	–	*	*
S01	Anz. Suchen	ja	96	R	–	–	+	+
S02	Erste betr. Rankingpos.	ja	*	*	–	–	80	R
S03	Letzte betr. Rankingpos.	ja	96	R	–	–	*	*
S04	Suchdauer	nein	108	R	40	R	96	R
S05-log	Zeit zum ersten richtig rel. bew. Dok.	nein	108	K	*	*	+	+
S05	Zeit zum ersten richtig rel. bew. Dok.	nein	108	R	40	R	+	+
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	ja	96	R	–	–	+	+
F04	Liefert die Suchmaschine genügend Information?	ja	*	*	–	–	80	R
F05	Ist die Suchmaschine präzise?	nein	+	+	*	*	96	R
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	ja	96	R	–	–	80	R
F15	Es war einfach, zu dem Thema zu suchen.	nein	*	*	40	R	*	*

[a] Nur einmal robuste Analyse.

[b] Nur einmal klassische Analyse.

Fortsetzung auf nächster Seite

Tab. C.16 (Fortsetzung)

			Aufgabe 1				Aufgabe 2	
			SP_A		SP_B		SP_A	
ID	**Beschreibung**	**topic**	**n**	**V**	**n**	**V**	**n**	**V**
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	nein	*	*	40	K/R	96	R
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	ja	96	R	–	–	+	+
F21	Ich hatte genügend Zeit, um eine effektive Suche durchzuführen.	ja	96	R	–	–	80	R
F22	Ich bin mit den Suchergebnissen zufrieden.	ja	+	+	–	–	80	R
F23	Ich bin mit meiner Suchleistung zufrieden.	ja	96	K/R[a]	–	–	*	*
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	ja	+	+	–	–	80	R
SK03-F	Benutzerfreundlichkeit	ja	+	+	–	–	80	K/R
SK06-F	Eigenleistung	nein	108	K	–	–	96	R
SK11-F	Eigenleistung	ja	96	K	–	–	+	+
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	nein	○	○	○	○	96	R
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	nein	○	○	○	○	96	R

[a] Nur einmal robuste Analyse.
[b] Nur einmal klassische Analyse.

C.3.2. Mittelwerte der Interaktionen

Da im Rahmen des zweiten Experiments lediglich für das Benutzerleistungsmaß M02 eine signifikante Interaktion zwischen Systemleistung und Erwartungshaltung nachgewiesen werden kann, enthalten die Tabellen im Hauptteil der Arbeit aus Gründen der Übersichtlichkeit ausschließlich die nach Erwartungshaltung und Systemgüte getrennt berechneten Gruppenmittelwerte. Der Vollständigkeit halber sind diese Tabellen im folgenden Abschnitt noch einmal, um die Gruppenmittelwerte der vier unterschiedlichen Untersuchungsgruppen ergänzt, aufgeführt.

Tab. C.17.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung für $A1$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System		Erwartung		Interaktion			
ID	**Beschreibung**	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$	$\mathbf{I_{G,H}}$	$\mathbf{I_{G,N}}$	$\mathbf{I_{S,H}}$	$\mathbf{I_{S,N}}$
M05	Anz. aufg. irrel. Dok.	**2,48**[a]	**5,9**	4,63	3,75	2,5	2,46	6,75	5,04
M08	Anz. falsch rel. bew. Dok.	**0,81**[a]	**1,9**	1,59	1,12	1	0,62	2,17	1,62
M15	Anz. richtig irrel. bew. Dok.	**1,69**	**3,78**[a]	2,8	2,67	1,74	1,63	3,85	3,7
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	**0,71**	**0,81**[a]	0,76	0,76	0,7	0,71	0,81	0,8
V01	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	**0,21**[a]	**0,4**	0,3	0,31	0,22	0,19	0,37	0,42
V02	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	**0,79**[a]	**0,61**	0,7	0,71	0,77	0,81	0,62	0,6
V05	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	**0,25**	**0,15**[a]	0,2	0,19	0,24	0,25	0,16	0,13
V06	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	**0,57**	**0,32**[a]	0,49	0,4	0,57	0,57	0,4	0,23
V08	Anz. falsch rel. bew. Dok. / Anz. aufg. Dok.	**0,07**[a]	**0,15**	0,13	0,09	0,07	0,06	0,18	0,12
V09	Anz. falsch rel. bew. Dok. / Anz. rel. bew. Dok.	**0,12**[a]	**0,24**	0,19	0,17	0,1	0,13	0,28	0,2
V10	Anz. falsch rel. bew. Dok. / Anz. richtig rel. bew. Dok.	**0,15**[a]	**0,4**	0,3	0,25	0,14	0,15	0,45	0,34
V14	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	**0,15**	**0,24**[a]	0,18	0,21	0,16	0,13	0,2	0,28
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	**0,43**	**0,69**[a]	0,52	0,6	0,45	0,41	0,59	0,79
V19	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)	**0,9**[a]	**0,75**	0,82	0,83	0,9	0,9	0,74	0,75
V21	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. bew. Dok. (erste Suche)	**0,9**[a]	**0,81**	0,84	0,88	0,91	0,89	0,76	0,86
V25	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. bew. Dok. (letzte Suche)	**0,9**[a]	**0,79**	0,85	0,84	0,9	0,89	0,8	0,78
V28/PCP	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.	**0,56**[a]	**0,47**	0,51	0,52	0,54	0,58	0,47	0,46
V31/BP	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.	**0,91**[a]	**0,78**	0,83	0,87	0,91	0,9	0,74	0,83

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Tab. C.18.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung für $A1$ in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. (SP_B Aufgabe 1)

		System		Erwartung		Interaktion			
ID	**Beschreibung**	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$	$\mathbf{I_{G,H}}$	$\mathbf{I_{G,N}}$	$\mathbf{I_{S,H}}$	$\mathbf{I_{S,N}}$
V02	Anz. aufg. rel. Dok.	**0,77**[a]	**0,59**	0,70	0,67	0,75	0,79	0,64	0,54

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

C.3.3. Teststatistiken der Varianzanalysen

Die in diesem Abschnitt dargestellten Tabellen enthalten weiterführende Informationen zu den im Rahmen des zweiten Experiments durchgeführten Varianzanalysen. Neben der jeweiligen Stichprobengröße, geben sie Auskunft darüber, ob der zugrunde liegende Datensatz topicbalanciert ist (topic), eine klassische (K) oder robuste (R) Varianzanalyse durchgeführt wurde oder es sich um einen eindeutigen (E) oder tendenziellen (T) Effekt handelt. In den mit *test* überschriebenen Spalten wird im Fall der klassischen Analyse der erreichte F-Wert angegeben, für robuste Analysen hingegen der entsprechende Testwert. Die mit *p* überschriebenen Spalten enthalten das zugehörige Signifikanzniveau. Die Freiheitsgrade der klassischen Analyse betragen bei allen Variablen der ersten Aufgabe $df = 1/92$ und bei der zweiten Aufgabe mit Ausnahme der Variable V02 ($df = 1/88$) $df = 1/72$.

Tab. C.23.: Teststatistik der Varianzanalyse zur Untersuchung des Einflusses von System, Erwartung und deren Wechselwirkung auf Benutzerleistung und -zufriedenheit für $A1$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**topic**	**V**	**Q**	**test**	**p**	**test**	**p**	**test**	**p**
M05	Anz. aufg. irrel. Dok.	96	ja	R	E	11,32	**0,003**	0,32	0,576	0,24	0,628
M08	Anz. falsch rel. bew. Dok.	96	ja	R	E	12,27	**0,001**	1,36	0,249	0,55	0,462
M15	Anz. richtig irrel. bew. Dok.	108	nein	R	E	9,8	**0,004**	0,14	0,71	1	0,323
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	88	ja	R	T	8,12	**0,007**	0,01	0,92	0,27	0,605
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	96	ja	K	E	51,94	$\mathbf{2 \cdot 10^{-10}}$	0,15	0,698	1,86	0,176
V02	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	108	nein	R	E	41,03	**0,001**	0,16	0,692	1,24	0,271
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	ja	R	E	12,24	**0,001**	0,1	0,748	0,13	0,722
V06	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	96	ja	R	E	24,96	**0,001**	1,23	0,274	2,99	0,091
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	ja	R	E	16,18	**0,001**	2,96	0,091	0,62	0,433
V09	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	96	ja	R	E	20,86	**0,001**	1,99	0,164	3,25	0,078
V10	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. richtig rel. bew. Dok.}}$	96	ja	R	E	23,48	**0,001**	4,45	0,04	1,94	0,17
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	108	nein	R	E	17,79	**0,001**	0,98	0,328	4,06	0,049
V17	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	96	ja	R	E	21,75	**0,001**	0,82	0,369	3,82	0,056

[a] Entspricht auch der Skala SK13-M.

[b] Entspricht auch der Skala SK18-M.

Fortsetzung auf nächster Seite

Tab. C.23 (Fortsetzung)

ID	Beschreibung	n	topic	V	Q	System test	System p	Erwartung test	Erwartung p	Interaktion test	Interaktion p
V21	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. bew. Dok. (erste Suche)	88	ja	R	E	10,08	**0,003**	2,41	0,127	3,55	0,066
V19	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)	88	ja	R	E	12,99	**0,001**	0,002	0,963	0,004	0,951
V25	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. bew. Dok. (letzte Suche)	88	ja	R	T	9,79	**0,003**	0,05	0,826	0,002	0,968
V28/PCP	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.	96	ja	R	T	8,63	**0,005**	0,01	0,91	1,77	0,19
V31/BP	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.	96	ja	R	E	19,65	**0,001**	2,63	0,111	2,07	0,157
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	96	ja	R	E	3,51	0,067	12,98	**0,001**	0,02	0,886
F05	Ist die Suchmaschine präzise?	108	nein	R	E	4,38	0,042	16,39	**0,001**	1,58	0,215
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	96	ja	R	E	0,74	0,394	14,48	**0,001**	0,38	0,542
F08	Ist die Suchmaschine benutzerfreundlich?	96	ja	R	E	0,14	0,714	18,51	**0,001**	1,83	0,182
F09	Ist die Suchmaschine einfach zu bedienen?	96	ja	R	E	0,22	0,641	10,86	**0,002**	0,89	0,352
F11	Liefert die Suchmaschine aktuelle Information?	96	ja	R	E	3,59	0,064	12,07	**0,002**	0,4	0,531
F12	Ist die Suchmaschine erfolgreich?	96	ja	R	E	1,2	0,278	13,35	**0,001**	3,34	0,074
F13	Sind Sie mit der Suchmaschine zufrieden?	96	ja	R	E	0,68	0,414	14,77	**0,001**	0,08	0,785
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	96	ja	R	E	0,05	0,827	12,43	**0,001**	0,05	0,827
F22	Ich bin mit den Suchergebnissen zufrieden.	96	ja	R	T	0,01	0,921	8,56	**0,005**	0,09	0,764
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	96	ja	R	E	0,75	0,391	9,43	**0,004**	0,01	0,938
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	96	ja	R	E	0,42	0,522	14,97	**0,001**	0,1	0,749
SK01-M	Genauigkeit	96	ja	R	E	5,04	0,03	14,76	**0,001**	0,3	0,586
SK01-F	Genauigkeit	96	ja	R	E	1,03	0,316	12,68	**0,001**	0,05	0,823
SK02-M	Inhalt	96	ja	R	E	0,44	0,508	9	**0,004**	0,44	0,508
SK02-F	Inhalt	96	ja	R	E	0,16	0,69	8,7	**0,005**	1,86	0,179
SK03-M	Benutzerfreundlichkeit	96	ja	R	E	1,13	0,294	16,45	**0,001**	1,72	0,196
SK03-F	Benutzerfreundlichkeit	96	ja	R	E	0,4	0,53	15,18	**0,001**	0,83	0,367
SK04-M	Suche	96	ja	R	T	0,94	0,337	11,53	**0,002**	0,01	0,924
SK04-F	Suche	96	ja	R	T	0,74	0,395	6,48	**0,015**	0,51	0,48
SK06-M	Eigenleistung	96	ja	K	T	0,34	0,563	4,75	**0,032**	0,4	0,53
SK07[b]M	Benutzerfreundlichkeit	108	nein	R	E	0,21	0,648	6,41	**0,015**	0,12	0,732
SK07-F	Benutzerfreundlichkeit	96	ja	K	E	0,72	0,398	10,34	**0,002**	0,6	0,44

[a] Entspricht auch der Skala SK13-M.

[b] Entspricht auch der Skala SK18-M.

Fortsetzung auf nächster Seite

Tab. C.23 (Fortsetzung)

ID	Beschreibung	n	topic	V	Q	System test	System p	Erwartung test	Erwartung p	Interaktion test	Interaktion p
SK08-M	Suche	96	ja	R	E	0,86	0,358	13,63	**0,001**	2,52	0,119
SK08-F	Suche	96	ja	R	E	0,83	0,366	10,17	**0,003**	0,77	0,386
SK09-M	Benutzerfreundlichkeit	96	ja	R	E	0,97	0,329	15,45	**0,001**	0,61	0,44
SK09-F	Benutzerfreundlichkeit	96	ja	R	E	0,94	0,339	12,91	**0,001**	0,98	0,328
SK-C	Content (EUCS)	96	ja	R	E	0,79	0,377	7,15	**0,01**	0,16	0,687
SK-A	Accuracy (EUCS)	96	ja	R	E	2,04	0,159	23,62	**0,001**	0,33	0,57
SK-E[a]	Ease of Use (EUCS)	96	ja	R	E	0,16	0,69	13,12	**0,001**	0,16	0,69
SK-T	Timeliness (EUCS)	108	nein	R	E	0,64	0,427	10,28	**0,003**	0,36	0,551
SK-K	Kriteriumsskala	96	ja	R	E	0,85	0,362	12,94	**0,001**	0,34	0,561
SK-E-88	EUCS-Skala-1988	96	ja	R	E	0,08	0,775	14,02	**0,001**	0,01	0,935
SK-E-09	EUCS-Skala-2009	96	ja	R	E	1,32	0,255	14,01	**0,001**	0,66	0,419
SK-Z-09	Zusatzskala-2009	96	ja	K	E	$1 \cdot 10^{-4}$	0,992	9,64	**0,003**	0,02	0,877
SK-G-09	Gesamtskala-2009	96	ja	R	E	0,83	0,365	15,6	**0,001**	1,24	0,27

[a] Entspricht auch der Skala SK13-M.
[b] Entspricht auch der Skala SK18-M.

Tab. C.19.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System 1, System 2 und Erwartung auf die Benutzerleistung für $A2$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System 1		System 2		Erwartung		Interaktion					
ID	Beschreibung	$S1_G$	$S1_S$	$S2_G$	$S2_S$	E_H	E_N	$I_{G,G,H}$	$I_{G,G,N}$	$I_{G,S,H}$	$I_{G,S,N}$	$I_{S,G,H}$	$I_{S,G,N}$	$I_{S,S,H}$	$I_{S,S,N}$
M03[b]	Anz. aufg. Dok. (erste Suche)	13,42	11,67	**15,84**	**9,25**	11,36	13,73	15,92	16,17	12,58	9	9,42	21,83	7,5	7,92
		13,96	11,69	15,40	10,25	13,00	12,65	16,50	16,67	13,67	9,00	11,42	17,00	10,42	7,92
M05	Anz. aufg. irrel. Dok.	4,58	6,43	**4,1**[a]	**6,9**	4,52	6,48	3	3,7	6,6	5	2,8	6,9	5,7	10,3
M07	Anz. falsch irrel. bew. Dok.	2,88	2,4	**3,13**	**2,15**[a]	2,13	3,15	3,4	3,5	2	2,6	1,8	3,8	1,3	2,7
M08	Anz. falsch rel. bew. Dok.	1,88	1,85	**1,2**[a]	**2,53**	1,65	2,08	0,7	2	2,5	2,3	0,9	1,2	2,5	2,8
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	3,38	3,08	**4,1**[a]	**2,35**	3,48	2,98	4	4,5	2,9	2,1	4,2	3,7	2,8	1,6
M12	Anz. rel. bew. Dok. (erste Suche)	8,43	7,83	**10,58**[a]	**5,68**	7,57	8,68	11,4	9,8	6,5	6	6,4	14,7	6	4,2
M15	Anz. richtig irrel. bew. Dok.	3,52	4,06	**2,56**	**5,02**[a]	3,54	4,04	2,58	2,75	5,08	3,67	1,83	3,08	4,67	6,67
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	3,04	2,48	**3,88**[a]	**1,65**	2,98	2,54	4,08	4,25	2,25	1,58	3,75	3,42	1,83	0,92
M18	Anz. richtig rel. bew. Dok. (erste Suche)	6,63	6,04	**8,46**[a]	**4,21**	6,06	6,61	9,08	8,25	5	4,17	6	10,5	4,17	3,5
V01	Anz. aufg. irrel. Dok.	0,28	0,31	**0,18**[a]	**0,42**	0,27	0,32	0,12	0,23	0,38	0,39	0,17	0,19	0,41	0,48
	Anz. aufg. Dok.														
V02	Anz. aufg. rel. Dok.	0,72	0,68	**0,8**[a]	**0,59**	0,71	0,68	0,84	0,79	0,6	0,63	0,8	0,77	0,61	0,52
	Anz. aufg. Dok.														
V06	Anz. falsch irrel. bew. Dok.	0,51	0,38	**0,57**	**0,33**[a]	0,42	0,48	0,61	0,68	0,38	0,38	0,38	0,61	0,29	0,26
	Anz. irrel. bew. Dok.														
V08	Anz. falsch rel. bew. Dok.	0,11	0,1	**0,06**[a]	**0,14**	0,11	0,1	0,03	0,1	0,18	0,13	0,05	0,07	0,16	0,11
	Anz. aufg. Dok.														
V09	Anz. falsch rel. bew. Dok.	0,16	0,18	**0,1**[a]	**0,24**	0,15	0,19	0,04	0,18	0,23	0,18	0,05	0,13	0,29	0,25
	Anz. rel. bew. Dok.														
V10	Anz. falsch rel. bew. Dok.	0,23	0,29	**0,11**[a]	**0,41**	0,24	0,29	0,05	0,18	0,39	0,3	0,08	0,14	0,42	0,52
	Anz. richtig rel. bew. Dok.														
V14	Anz. richtig irrel. bew. Dok.	0,19	0,2	**0,13**	**0,26**[a]	0,17	0,21	0,11	0,12	0,27	0,25	0,11	0,16	0,2	0,32
	Anz. aufg. Dok.														
V16	Anz. richtig irrel. bew. Dok.	1,25	1,7	**0,86**	**2,1**[a]	1,57	1,39	0,7	0,9	2,05	1,37	1,01	0,82	2,51	2,47
	Anz. falsch irrel. bew. Dok.														
V17	Anz. richtig irrel. bew. Dok.	0,47	0,53	**0,41**	**0,6**[a]	0,49	0,51	0,35	0,34	0,58	0,62	0,44	0,49	0,6	0,6
	Anz. irrel. bew. Dok.														
V28/PCP	Anz. richtig rel. bew. Dok.	0,53	0,52	**0,6**[a]	**0,45**	**0,56**[a]	**0,49**	0,65	0,59	0,4	0,48	0,66	0,51	0,53	0,37
	Anz. aufg. Dok.														
V31/BP	Anz. richtig rel. bew. Dok.	0,83	0,84	**0,9**[a]	**0,77**	0,85	0,82	0,96	0,83	0,75	0,78	0,92	0,9	0,78	0,76
	Anz. rel. bew. Dok.														
S01	Anz. Suchen	2,08	2,45	**1,78**[a]	**2,75**	2,15	2,38	2,1	1,6	1,8	2,8	2,1	1,3	2,6	3,8
S05	Zeit zum ersten richtig rel. bew. Dok.	124,29	116,35	**94,77**[a]	**145,88**	130,37	110,27	117,33	79,92	171,75	128,17	105,83	76	126,58	157
S05-log	Zeit zum ersten richtig rel. bew. Dok.	4,59	4,51	**4,34**[a]	**4,76**	4,58	4,52	4,55	4,21	4,9	4,68	4,34	4,27	4,53	4,91

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.
[b] Für M03 wird auch die Wechselwirkung zwischen der Systemleistung des in Aufgabe 2 verwendeten Systems und der Erwartungshaltung signifikant.

Tab. C.20.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerzufriedenheit für $A1$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	3,67	3,96	**4,13**[c]	**3,5**	4,04	3,29	4,21	3,71
F05	Ist die Suchmaschine präzise?	3,47	3,11	**3,63**[c]	**2,95**	3,74	3,19	3,52	2,7
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	3,54	3,33	**3,85**[c]	**3,02**	4,08	3	3,62	3,04
F08	Ist die Suchmaschine benutzerfreundlich?	3,63	3,78	**4,21**[c]	**3,19**	4,04	3,21	4,38	3,17
F09	Ist die Suchmaschine einfach zu bedienen?	4,5	4,54	**4,73**[c]	**4,31**	4,67	4,33	4,79	4,29
F11	Liefert die Suchmaschine aktuelle Information?	3,52	3,87	**4,02**[c]	**3,37**	3,92	3,12	4,12	3,62
F12	Ist die Suchmaschine erfolgreich?	3,77	3,92	**4,13**[c]	**3,56**	3,96	3,58	4,29	3,54
F13	Sind Sie mit der Suchmaschine zufrieden?	3,8	3,57	**4,09**[c]	**3,28**	4,21	3,38	3,96	3,17
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	4,73	4,79	**5,21**[c]	**4,31**	5,17	4,29	5,25	4,33
F22	Ich bin mit den Suchergebnissen zufrieden.	4,64	4,67	**5,11**[c]	**4,21**	5,17	4,12	5,04	4,29
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	4,71	4,44	**5,15**[c]	**4**	5,33	4,08	4,96	3,92
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	4,17	4,05	**4,84**[c]	**3,38**	4,96	3,38	4,71	3,38
SK01-M	Genauigkeit	0,63	0,53	**0,66**[c]	**0,5**	0,7	0,55	0,62	0,44
SK01-F	Genauigkeit	0,09	-0,005	**0,36**[c]	**-0,28**	0,38	-0,21	0,34	-0,35
SK02-M	Inhalt	0,62	0,64	**0,7**[c]	**0,56**	0,67	0,56	0,72	0,55
SK02-F	Inhalt	-0,12	-0,02	**0,23**[c]	**-0,37**	0,08	-0,32	0,37	-0,42
SK03-M	Benutzerfreundlichkeit	0,61	0,57	**0,71**[c]	**0,48**	0,7	0,52	0,72	0,43
SK03-F	Benutzerfreundlichkeit	0,02	-0,02	**0,47**[c]	**-0,47**	0,44	-0,39	0,5	-0,54
SK04-M	Suche	0,65	0,69	**0,73**[c]	**0,62**	0,71	0,59	0,74	0,64
SK04-F	Suche	0,02	0,13	**0,35**[c]	**-0,19**	0,22	-0,18	0,47	-0,2
SK06-M	Eigenleistung	0,47	0,5	**0,53**[c]	**0,43**	0,5	0,43	0,56	0,43
SK07-M[a]	Benutzerfreundlichkeit	0,62	0,61	**0,68**[c]	**0,55**	0,67	0,57	0,68	0,54
SK07-F	Benutzerfreundlichkeit	-0,05	-0,23	**0,19**[c]	**-0,47**	0,2	-0,3	0,18	-0,64
SK08-M	Suche	0,65	0,62	**0,71**[c]	**0,56**	0,7	0,6	0,71	0,53
SK08-F	Suche	-0,05	0,09	**0,35**[c]	**-0,31**	0,23	-0,33	0,46	-0,29
SK09-M	Benutzerfreundlichkeit	0,62	0,58	**0,7**[c]	**0,5**	0,71	0,52	0,69	0,47
SK09-F	Benutzerfreundlichkeit	0,02	-0,1	**0,4**[c]	**-0,47**	0,37	-0,32	0,42	-0,62
SK-A	Accuracy (EUCS)	0,6	0,55	**0,69**[c]	**0,45**	0,7	0,49	0,67	0,42
SK-C	Content (EUCS)	0,62	0,66	**0,7**[c]	**0,57**	0,67	0,56	0,72	0,59
SK-E[b]	Ease of Use (EUCS)	0,76	0,81	**0,88**[c]	**0,69**	0,85	0,67	0,9	0,71
SK-T	Timeliness (EUCS)	0,71	0,73	**0,79**[c]	**0,65**	0,79	0,63	0,79	0,67
SK-K	Kriteriumsskala	0,7	0,74	**0,81**[c]	**0,64**	0,79	0,61	0,82	0,66
SK-E-88	EUCS-Skala-1988	0,67	0,69	**0,75**[c]	**0,61**	0,75	0,59	0,74	0,63
SK-E-09	EUCS-Skala-2009	0,64	0,59	**0,7**[c]	**0,53**	0,71	0,56	0,68	0,5
SK-Z-09	Zusatzskala-2009	0,61	0,61	**0,66**[c]	**0,56**	0,65	0,56	0,66	0,55
SK-G-09	Gesamtskala-2009	0,64	0,63	**0,69**[c]	**0,57**	0,69	0,59	0,69	0,56

[a] Entspricht auch der Skala SK18-M.
[b] Entspricht auch der Skala SK13-M.
[c] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit.

Tab. C.21.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerzufriedenheit für $A2$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System 1		System 2		Erwartung		Interaktion						
ID	Beschreibung	$S1_G$	$S1_S$	$S2_G$	$S2_S$	E_H	E_N	$I_{G,G,H}$	$I_{G,G,N}$	$I_{G,S,H}$	$I_{G,S,N}$	$I_{S,G,H}$	$I_{S,G,N}$	$I_{S,S,H}$	$I_{S,S,N}$
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	3,63	3,55	**3,97**[c]	**3,2**	3,68	3,5	3,9	4	3,6	3	3,8	4,2	3,4	2,8
F12	Ist die Suchmaschine erfolgreich?	3,65	3,85	**4,02**[c]	**3,48**	3,88	3,63	4,4	3,8	3,3	3,1	3,7	4,2	4,1	3,4
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	4,7	4,95	**5,3**[c]	**4,35**	5,05	4,6	5,2	5,2	4,4	4	5,5	5,3	5,1	3,9
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	4,65	4,7	**5,1**[c]	**4,25**	4,85	4,5	5,3	4,9	4,5	3,9	5	5,2	4,6	4
SK04-F	Suche	-0,18	0,04	**0,29**[c]	**-0,42**	0,12	-0,25	0,27	-0,12	-0,18	-0,68	0,4	0,59	-0,03	-0,78
SK05-F	Aufgabe	-0,05	-0,02	**0,17**[c]	**-0,25**	-0,18	0,11	-0,08	0,28	-0,29	-0,12	-0,12	0,61	-0,23	-0,34
SK07-M[a]	Benutzerfreundlichkeit	0,62	0,64	**0,69**[c]	**0,56**	0,6	0,65	0,67	0,67	0,52	0,62	0,69	0,74	0,53	0,58
SK08-F	Suche	-0,02	-0,13	**0,26**[c]	**-0,41**	0,04	-0,19	0,37	0,29	-0,28	-0,47	0,04	0,32	0,02	-0,9
SK11-M[b]	Eigenleistung	0,57	0,56	**0,61**[c]	**0,51**	0,57	0,56	0,61	0,63	0,52	0,52	0,54	0,67	0,6	0,41
SK11-F	Eigenleistung	0,09	0,06	**0,36**[c]	**-0,22**	-0,13	0,28	0,12	0,57	-0,33	-0,01	-0,1	0,85	-0,22	-0,3
SK-A	Accuracy (EUCS)	0,56	0,61	**0,67**[c]	**0,5**	0,62	0,55	0,68	0,6	0,5	0,46	0,71	0,68	0,59	0,46
SK-K	Kriteriumsskala	0,64	0,63	**0,73**[c]	**0,54**	0,69	0,58	0,84	0,65	0,6	0,48	0,69	0,74	0,61	0,46
SK-Z-09	Zusatzskala-2009	0,63	0,64	**0,68**[c]	**0,59**	0,62	0,65	0,7	0,68	0,54	0,59	0,63	0,72	0,62	0,6
SK-G-09	Gesamtskala-2009	0,63	0,65	**0,7**[c]	**0,57**	0,64	0,64	0,73	0,66	0,55	0,56	0,66	0,75	0,6	0,59

[a] Entspricht auch der Skala SK18-M.
[b] Entspricht auch den Skalen SK15-M und SK19-M.
[c] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit.

Tab. C.22.: Teststatistik der Varianzanalyse zur Untersuchung des Einflusses von System, Erwartung und deren Wechselwirkung auf Benutzerleistung und -zufriedenheit für $A1$ in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

					System		Erwartung		Interaktion	
ID	n	topic	V	Q	test	p	test	p	test	p
V02	40	nein	R/K	E	19,12	$9{,}6 \cdot 10^{-5}$	0,47	0,498	2,76	0,105

Tab. C.24.: Teststatistik der Varianzanalyse zur Untersuchung des Einflusses von System 1, System 2, Erwartung und deren Wechselwirkung auf Benutzerleistung und -zufriedenheit für $A2$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

					S1		S2		E		S1 - E		S2 - E		S1 - S2		S1 - S2 - E	
ID	**n**	**topic**	**V**	**Q**	**test**	**p**	**test**	**p**	**test**	**p**	**test**	**p**	**test**	**p**	**test**	**p**	**test**	**p**
M03	96	nein	R	E	0,68	0,416	7,77	**0,008**	1,13	0,3	1,07	0,307	5,09	**0,03**	0,63	0,431	0,07	0,792
					0,63	0,43	4,56	**0,038**	0,05	0,83	0,11	0,745	5,41	**0,025**	0,07	0,786	0,02	0,892
M05	80	ja	R	E	4,4	0,043	21,75	$\mathbf{1 \cdot 10^{-4}}$	4,4	0,043	4,4	0,043	$1 \cdot 10^{-31}$	0,999	0,22	0,644	1,96	0,171
M07	80	ja	R	E	1,33	0,26	17,2	$\mathbf{3 \cdot 10^{-4}}$	3,4	0,076	1,91	0,178	0,05	0,82	0,21	0,649	2,6	0,118
M08	80	ja	R	E	0,09	0,769	13,7	**0,002**	0,81	0,39	0,01	0,922	0,81	0,383	0,49	0,495	1,69	0,213
M11	80	ja	R	E	0,83	0,368	13,96	**0,001**	2,49	0,13	0,83	0,368	3,04	0,091	1,99	0,168	0,01	0,935
M12	80	ja	R	E	0,07	0,799	8,37	**0,009**	0,22	0,65	0,13	0,722	5,17	0,034	0,87	0,364	2,09	0,163
M15	96	nein	R	E	0,62	0,434	11,5	**0,002**	0,62	0,44	0,95	0,335	0,78	0,383	0,36	0,55	0,05	0,818
M17	96	nein	R	E	1,43	0,239	26,33	$\mathbf{1 \cdot 10^{-4}}$	1,78	0,19	0,4	0,531	2,18	0,148	0,6	0,444	0,04	0,834
M18	96	nein	R	E	0,01	0,917	10,98	**0,002**	0,28	0,61	0,04	0,835	5,12	0,029	0,8	0,377	0,8	0,377
V01	80	ja	K	E	1,96	0,166	97,62	$\mathbf{1 \cdot 10^{-14}}$	4,68	0,034	0,24	0,626	0,21	0,651	1,19	0,278	2,08	0,154
V02	96	nein	K	E	2,4	0,125	69,18	$\mathbf{1 \cdot 10^{-12}}$	2,08	0,153	1,11	0,296	0,03	0,854	0,1	0,75	1,84	0,179
V06	80	ja	R	E	6,29	0,019	18,61	$\mathbf{2 \cdot 10^{-4}}$	4,97	0,035	$4 \cdot 10^{-4}$	0,985	2,11	0,159	1,41	0,246	1,07	0,311
V08	80	ja	R	E	0,2	0,661	18,6	$\mathbf{2 \cdot 10^{-4}}$	0,1	0,76	0,74	0,394	3,9	0,056	0,48	0,493	0,03	0,856
V09	80	ja	R	E	0,36	0,554	16,89	$\mathbf{3 \cdot 10^{-4}}$	1,24	0,28	0,63	0,434	7,59	0,01	1,18	0,287	0,06	0,81
V10	80	ja	R	E	0,18	0,676	25,74	$\mathbf{1 \cdot 10^{-4}}$	1,12	0,31	0,27	0,609	1,66	0,21	0,35	0,558	0,01	0,913
V14	96	nein	R	E	0,47	0,498	21	$\mathbf{1 \cdot 10^{-4}}$	3,59	0,066	4,88	0,033	0,19	0,666	0,11	0,742	1,7	0,201
V16	80	ja	R	E	6,79	0,015	33,98	$\mathbf{1 \cdot 10^{-4}}$	1,08	0,31	0,62	0,439	0,46	0,506	0,75	0,394	0,31	0,585
V17	80	ja	R	E	2,17	0,152	22,51	$\mathbf{1 \cdot 10^{-4}}$	0,14	0,72	0,4	0,533	0,11	0,745	0,61	0,441	0,28	0,605
V28/PCP	80	ja	K	E	0,13	0,718	19,16	$\mathbf{4 \cdot 10^{-5}}$	4,2	**0,044**	5,69	0,02	0,77	0,383	0,51	0,478	1,02	0,317
V31/BP	80	ja	R	E	0,9	0,354	25,29	$\mathbf{1 \cdot 10^{-4}}$	1,69	0,21	2,02	0,169	1,34	0,259	0,002	0,965	0,74	0,398
S01	80	ja	R	E	5,51	0,029	11,92	**0,003**	1,54	0,23	0,02	0,892	5,51	0,029	3,22	0,088	0,48	0,499
S05	96	nein	R	E	0,57	0,455	10,41	**0,003**	2,38	0,14	1	0,325	0,57	0,455	0,08	0,78	0,48	0,492
S05-log	96	nein	R	T	0,66	0,424	12,2	**0,002**	0,67	0,42	2,91	0,097	1,84	0,184	0,01	0,942	0,8	0,379
F03	80	ja	R	E	0	0,999	19,51	$\mathbf{2 \cdot 10^{-4}}$	2,17	0,16	0,14	0,717	6,64	0,017	0,14	0,717	0	0,999
F12	80	ja	R	E	1,17	0,289	6,38	**0,018**	1,17	0,29	1,17	0,289	1,17	0,289	1,17	0,289	1,17	0,289
F18	80	ja	R	E	1,59	0,217	10,37	**0,003**	3,31	0,079	1,59	0,217	3,31	0,079	0,49	0,49	0,49	0,49
F19	80	ja	R	T	0,02	0,902	9,81	**0,004**	1,27	0,27	0,39	0,536	0,77	0,388	0,02	0,902	0,14	0,71
SK04-F	80	ja	K	T	1	0,321	9,93	**0,002**	2,6	0,111	0,13	0,724	1,38	0,243	0,79	0,378	0,85	0,359
SK05-F	96	nein	R	T	$1 \cdot 10^{-4}$	0,994	4,71	**0,036**	2,25	0,15	0,18	0,672	2,53	0,12	1,23	0,275	0,21	0,653
SK07-M	96	nein	R	T	0,17	0,685	9,59	**0,004**	2,49	0,13	0,003	0,954	0,17	0,685	0,28	0,603	0,58	0,453
SK08-F	80	ja	R	E	0,18	0,676	10,42	**0,003**	5,04	0,033	0,17	0,685	2,14	0,154	0,002	0,962	0,61	0,443
SK11-M	80	ja	K	T	0,15	0,702	5,33	**0,024**	0,04	0,848	0,33	0,566	3,7	0,058	0	1	3	0,088
SK11-F	80	ja	R	T	0,04	0,839	13,84	**0,001**	2,48	0,13	0,31	0,583	0,88	0,358	0,75	0,395	1,96	0,174
SK-A	80	ja	K	T	0,68	0,412	7,18	**0,009**	1,29	0,261	0,04	0,837	0,04	0,837	0,01	0,918	0,27	0,608
SK-K	80	ja	R	E	0,04	0,848	12,33	**0,002**	6,43	0,019	1,37	0,254	1,86	0,186	0,04	0,848	0,61	0,444
SK-Z-09	80	ja	K	T	0,16	0,692	5,95	**0,017**	0,43	0,514	0,14	0,708	0,1	0,756	0,46	0,501	1,32	0,254
SK-G-09	80	ja	R	T	0,7	0,411	6,66	**0,016**	0,001	0,98	0,4	0,532	0,14	0,715	0,13	0,722	0,36	0,555

C.4. Weitere Ergebnisse der Topiceffektanalyse

In diesem Anhang sind weitergehende Ergebnisse der Topiceffektanalyse von Experiment 2 zusammengefasst. Insbesondere können die Ergebnisse der Varianzanalysen von SP_A unter Ausschluss der im Rahmen des zweiten Experiments identifizierten, möglicherweise problematischen Fallgruppen betrachtet werden. Bevor diese Ergebnisse jedoch in Abschnitt C.4.2 wiedergegeben sind, stellt Abschnitt C.4.1 zunächst eine Übersicht derjenigen Variablen bereit, für die sich in keiner der untersuchten Stichproben und bei keiner der beiden Testaufgaben ein signifikanter Topiceffekt nachweisen lässt.

C.4.1. Variablen ohne signifikante Unterschiede

Die im Folgenden präsentierte Tabelle C.25 stellt eine Übersicht der 38 Variablen bereit, die in keiner der untersuchten Stichproben und bei keiner der beiden Aufgaben einen signifikanten Topiceffekt zeigen. Neben der Stichprobengröße gibt die Tabelle auch Auskunft darüber, ob eine klassische (K) oder robuste (R) Varianzanalyse durchgeführt wird. Eine Fußnote markiert darüber hinaus Fälle, in denen eine der beiden Analysevarianten die Ausnahme bleibt. Vereinzelt treten auch Fälle auf, in denen die Durchführung einer Varianzanalyse aufgrund zu geringer Fallzahlen nicht möglich ist (in der Tabelle durch einen waagerechten Strich symbolisiert). Die Erwartungsitems hingegen werden ausschließlich im Anschluss an die zweite Suchaufgabe erhoben und können deshalb prinzipiell nicht für die erste Aufgabe ausgewertet werden. Auf diesen Umstand verweist in der Tabelle ein Kreis (○).

Tab. C.25.: Übersicht über Benutzerleistungs- und Zufriedenheitsvariablen ohne Topiceffekt. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde über alle fünf Stichproben hinweg oder einzeln pro Stichprobe eine klassische (K) oder robuste (R) Varianzanalyse durchgeführt.

		Aufgabe 1				**Aufgabe 2**			
		SP_A		**SP_B**		**SP_A**		**SP_B**	
ID	**Beschreibung**	**n**	**V**	**n**	**V**	**n**	**V**	**n**	**V**
M01	Anz. aufg. Dok.	96	R	32	K[a]/R	96	R	16	K/R[b]
M03	Anz. aufg. Dok. (erste Suche)	96	R	32	K/R	96	R	16	K/R[b]
M04	Anz. aufg. Dok. (letzte Suche)	96	R	32	K/R[b]	96	R	16	K R[b]
M06	Anz. aufg. rel. Dok.	96	R	32	K[a]/R	96	R	16	K
M15	Anz. richtig irrel. bew. Dok.	96	R	32	K[a]/R	96	R	16	K/R
M18	Anz. richtig rel. bew. Dok. (erste Suche)	96	R	32	K/R	96	R	16	K/R
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	96	R	32	K[a]/R	96	R	16	K
Z01	Durchschn. Betrachtungsz. aller Dok.	96	R	32	R	96	R	16	K/R
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	96	R	32	R	96	R	16	K[a]/R
Z04	Durchschn. Betrachtungsz. falsch rel. bew. Dok.	48	R	24	R	–	–	–	–
Z04-log	Durchschn. Betrachtungsz. falsch rel. bew. Dok.	48	R	24	R	–	–	–	–
Z06	Durchschn. Betrachtungsz. irrel. Dok.	72	R	32	R	64	R	–	–
Z06-log	Durchschn. Betrachtungsz. irrel. Dok.	72	R	32	K[a]/R	64	K	–	–
Z09-log	Durchschn. Betrachtungsz. richtig bew. Dok.	96	K	32	K	96	K	16	K/R
Z10	Durchschn. Betrachtungsz. richtig irrel. bew. Dok.	72	R	24	K[a]/R	64	R	–	–
Z10-log	Durchschn. Betrachtungsz. richtig irrel. bew. Dok.	72	K	24	K	64	K	–	–
V02	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	96	K	32	K	96	K	16	K
V03	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	96	R	32	K/R[b]	96	R	16	K/R[b]
V04	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	96	R	32	K/R[b]	96	R	16	K/R

[a] Nur einmal klassische Analyse.

[b] Nur einmal robuste Analyse.

Fortsetzung auf nächster Seite

Tab. C.25 (Fortsetzung)

		Aufgabe 1				Aufgabe 2			
		SP_A		SP_B		SP_A		SP_B	
ID	**Beschreibung**	**n**	**V**	**n**	**V**	**n**	**V**	**n**	**V**
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	K[a]/R	32	K	96	R	16	K
V30	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. falsch rel. bew. Dok.}}$	48	R	24	K/R	–	–	–	–
S04	Suchdauer	96	R	32	R	96	R	16	R
S05	Zeit zum ersten richtig rel. bew. Dok.	96	R	32	R	96	R	16	K/R
S05-log	Zeit zum ersten richtig rel. bew. Dok.	96	K/R	32	K/R	96	K	16	K
F05	Ist die Suchmaschine präzise?	96	R	32	R	96	R	16	K[a]/R
F15	Es war einfach, zu dem Thema zu suchen.	96	R	32	R	96	R	16	R
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	96	R	32	K[a]/R	96	R	16	K[a]/R
SK05-F	Aufgabe	96	R	32	K/R	96	R	16	K/R
SK06-F	Eigenleistung	96	K/R[b]	32	K	96	K	16	K
SK07-M	Benutzerfreundlichkeit	96	R	32	K/R	96	R	16	K/R[b]
SK-T	Timeliness (EUCS)	96	R	32	R	96	R	16	K/R
E01	Ich glaube, ich werde in zehn Minuten ... rel. Dok. finden.	○	○	○	○	64	R	16	R
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	○	○	○	○	96	R	16	K/R
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	○	○	○	○	96	R	16	R
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	○	○	○	○	96	R	16	K[a]/R
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	○	○	○	○	96	R	16	K/R
E06-M	Erwartungsskala	○	○	○	○	96	K/R[b]	16	K/R[b]

[a] Nur einmal klassische Analyse.
[b] Nur einmal robuste Analyse.

C.4.2. Ergebnisse der Topiceffektanalyse unter Ausschluss kritischer Fallgruppen

Die im Folgenden aufgeführten Tabellen enthalten die Ergebnisse der zusätzlich durchgeführten Varianzanalysen zur Bestimmung von Aufgabeneffekten unter Ausschluss bestimmter als problematisch eingeschätzter Fallgruppen. Im Gegensatz zu der in Abschnitt 6.4.6.1 aufgeführten Übersichtstabelle beinhalten diese Tabellen die Gruppenmittelwerte aller in Bezug auf die jeweilige Teilstichprobe signifikant werdenden Variablen. Die Tabellen unterscheiden sich in ihrem Aufbau nicht von den Tabellen der Hauptauswertung für die Stichprobe SP_A, sodass an dieser Stelle auf eine ausführliche Erläuterung verzichtet wird. Als zusätzliche Information gibt eine Fußnote an ob die Variable auch im Rahmen der Hauptauswertung signifikant ist. Bevor jedoch in den Tabellen C.27 bis C.40 die Ergebnisse der unter Ausschluss problematischer Fallgruppen durchgeführten Topiceffektanalysen angegeben werden, informiert Tabelle C.26 zunächst über im Kontext der Benutzerleistung signifikant werdende Topiceffekte in der Stichprobe SP_B, die aus Platzgründen nicht innerhalb der Arbeit behandelt werden.

Tab. C.26.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A1$ in SP_B. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M08	Anz. falsch rel. bew. Dok.	32	R	1/17	8,25	0,01	2,06	1,38[a]
V16	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.	24	R/K	1/8	11,17	0,0089	1,88[a]	1,46

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.

Tab. C.27.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A1$ in SP_A unter Ausschluss von SP_{MV}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M08	Anz. falsch rel. bew. Dok.	96	R	1/40	9,64	0,0035	1,75	1,02[a]
V05	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	96	R	1/51	8,07	0,0064	0,14[a]	0,23
V16	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.	56	R	1/23	9,42	0,0053	1,58[a]	0,75
V19	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)	80	R	1/36	5,16	0,029	0,79	0,90[a]

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.

Tab. C.28.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A1$ in SP_A unter Ausschluss von SP_{SB}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
V16	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.	32	R	1/11	8,52	0,013	1,79[a]	1,10

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.

Tab. C.29.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A1$ in SP_A unter Ausschluss von SP_{TD}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M05[b]	Anz. aufg. irrel. Dok.	88	R	1/38	9,15	0,0044	5,02	3,20[a]
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	88	R	1/48	14,71	0,00037	3,57[a]	2,25
Z03-log	Durchschn. Betrachtungsz. falsch irrel. bew. Dok.	56	K	1/54	6,69	0,012	3,01[a]	3,49
V06	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	72	R	1/40	12,87	0,00088	0,34[a]	0,55
V16	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. falsch irrel. bew. Dok.}}$	56	R	1/22	11,51	0,0026	1,62[a]	0,89
V17	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	72	R	1/37	7,19	0,011	0,62[a]	0,49

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.30.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A1$ in SP_A unter Ausschluss von SP_{UZ}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	80	R	1/41	6,70	0,013	3,55[a]	2,70
F19[b]	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	80	R	1/45	12,00	0,0012	5,08[a]	4,22
SK04-M[b]	Suche	80	R	1/40	10,14	0,0028	0,70[a]	0,59
SK08-M[b]	Suche	80	R	1/41	9,08	0,0044	0,69[a]	0,56

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.31.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A1$ in SP_A unter Ausschluss von SP_{TD}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F19[b]	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	88	R	1/52	10,11	0,0025	5,16[a]	4,32

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.32.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A1$ in SP_A unter Ausschluss von SP_{MV}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	96	R	1/56	8,75	0,0045	3,81[a]	3,06
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	96	R	1/48	9,27	0,0038	5,35[a]	4,79
F19[b]	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	96	R	1/57	14,74	0,00031	5,15[a]	4,29
SK04-M[b]	Suche	96	R	1/54	12,80	0,00074	0,73[a]	0,60
SK08-M[b]	Suche	96	R	1/51	8,38	0,0056	0,72[a]	0,59
SK11-M[c]	Eigenleistung	96	K/R	1/94	7,80	0,0063	0,56[a]	0,43
SK-C	Content (EUCS)	96	R	1/48	5,17	0,028	0,72[a]	0,58

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.
[c] Entspricht auch den Skalen SK15-M und SK19-M.

Tab. C.33.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A2$ in SP_A unter Ausschluss von SP_{UZ}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M08[b]	Anz. falsch rel. bew. Dok.	64	R	1/28	11,11	0,0024	3,16	1,34[a]
M10[b]	Anz. rel. bew. Dok.	64	R	1/25	7,89	0,0095	13,06[a]	7,84
B04[b]	Durchschn. Bew. rel. Dok.	64	R	1/28	12,27	0,0015	0,81[a]	0,64
V05[b]	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	64	R	1/31	8,29	0,0071	0,13[a]	0,27
V08[b]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	64	R	1/34	8,63	0,0059	0,14	0,072[a]
V11[b]	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	64	K/R	1/62	13,01	0,00062	0,33[a]	0,48
V12[b]	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	64	K/R	1/62	8,62	0,0046	0,67[a]	0,53
V13[b]	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$	64	K	1/62	8,76	0,0044	0,75[a]	0,65
V17	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	64	K/R	1/62	8,84	0,0042	0,65[a]	0,46
V29[b]	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	64	K/R	1/62	16,40	0,00014	0,81[a]	0,64

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.34.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A2$ in SP_A unter Ausschluss von SP_{MV}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M07[b]	Anz. falsch irrel. bew. Dok.	96	R	1/41	7,00	0,011	2,21[a]	3,92
M08[b]	Anz. falsch rel. bew. Dok.	96	R	1/43	10,39	0,0024	2,58	1,33[a]
M10[b]	Anz. rel. bew. Dok.	96	R	1/35	15,20	0,00041	12,21[a]	7,33
M13[b]	Anz. rel. bew. Dok. (letzte Suche)	96	R	1/42	7,72	0,0081	10,00[a]	6,48
M14	Anz. richtig bew. Dok.	64	R	1/36	9,05	0,0047	15,72[a]	10,28
M16[b]	Anz. richtig rel. bew. Dok.	96	R	1/38	11,43	0,0017	10,38[a]	6,85
M19	Anz. richtig rel. bew. Dok. (letzte Suche)	96	R	1/43	6,90	0,012	8,67[a]	5,17
B01[b]	Durchschn. Bew. irrel. Dok.	64	R	1/37	11,80	0,0015	0,45	0,22[a]
B04[b]	Durchschn. Bew. rel. Dok.	96	R	1/51	15,59	0,00024	0,81[a]	0,65
B05[b]	Durchschn. Bew. rel. Dok. (erste Suche)	80	R	1/43	17,30	0,00015	0,79[a]	0,63
B06[b]	Durchschn. Bew. rel. Dok. (letzte Suche)	80	R	1/39	12,60	0,001	0,83[a]	0,67
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	96	R	1/57	8,12	0,006	25,54[a]	31,64
Z11-log[b]	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	96	K	1/94	7,78	0,0064	2,90[a]	3,29
V05[b]	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	96	R	1/43	11,06	0,0018	0,13[a]	0,24
V07[b]	Anz. falsch irrel. bew. Dok. / Anz. richtig irrel. bew. Dok.	64	R	1/24	9,39	0,0052	0,61[a]	1,72
V08[b]	Anz. falsch rel. bew. Dok. / Anz. aufg. Dok.	96	R	1/54	13,80	0,00048	0,15	0,07[a]
V10	Anz. falsch rel. bew. Dok. / Anz. richtig rel. bew. Dok.	96	R	1/48	6,05	0,017	0,36	0,18[a]
V11[b]	Anz. irrel. bew. Dok. / Anz. aufg. Dok.	96	R/K	1/45	17,46	0,00013	0,32[a]	0,48
V12[b]	Anz. rel. bew. Dok. / Anz. aufg. Dok.	96	K/R	1/94	15,29	0,00017	0,69[a]	0,54
V13[b]	Anz. richtig bew. Dok. / Anz. aufg. Dok.	64	K	1/62	16,02	0,00017	0,77[a]	0,65
V29[b]	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	96	R	1/50	21,27	$2{,}8 \cdot 10^{-05}$	0,81[a]	0,63
V32/BR[b]	Anz. richtig rel. bew. Dok. / Anz. rel. Dok. im Korpus	96	R	1/37	12,59	0,0011	0,15[a]	0,091
V33[b]	Anz. richtig rel. bew. Dok. / Anz. zurückgeg. rel. Dok.	96	R	1/47	8,02	0,0068	0,19[a]	0,13

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.

[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.35.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A2$ in SP_A unter Ausschluss von SP_{SB}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M10[b]	Anz. rel. bew. Dok.	32	R/K	1/15	10,46	0,0053	14,38[a]	8,38
B04[b]	Durchschn. Bew. rel. Dok.	32	R/K	1/17	14,51	0,0014	0,83[a]	0,61
V05[b]	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	32	K/R	1/30	14,10	0,00074	0,13[a]	0,28
V12[b]	Anz. rel. bew. Dok. / Anz. aufg. Dok.	32	K/R	1/30	10,42	0,003	0,69[a]	0,49
V29[b]	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	32	K/R	1/30	15,67	0,00043	0,83[a]	0,64

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.36.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung für $A2$ in SP_A unter Ausschluss von SP_{TD}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
M07[b]	Anz. falsch irrel. bew. Dok.	80	R	1/35	11,17	0,002	2,33[a]	3,77
M10[b]	Anz. rel. bew. Dok.	80	R	1/35	9,55	0,0039	11,40[a]	8,12
B04[b]	Durchschn. Bew. rel. Dok.	80	R/K	1/37	16,91	$2 \cdot 10^{-04}$	0,82[a]	0,66
B05[b]	Durchschn. Bew. rel. Dok. (erste Suche)	64	R	1/37	13,81	0,00065	0,79[a]	0,63
B06[b]	Durchschn. Bew. rel. Dok. (letzte Suche)	64	R	1/33	14,10	0,00067	0,83[a]	0,65
V05[b]	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	80	R	1/42	19,87	$5{,}9 \cdot 10^{-05}$	0,11[a]	0,23
V07[b]	Anz. falsch irrel. bew. Dok. / Anz. richtig irrel. bew. Dok.	48	R	1/22	12,59	0,0018	0,47[a]	1,58
V11[b]	Anz. irrel. bew. Dok. / Anz. aufg. Dok.	80	R/K	1/36	17,31	0,00019	0,29[a]	0,45
V12[b]	Anz. rel. bew. Dok. / Anz. aufg. Dok.	80	R/K	1/37	12,93	0,00093	0,69[a]	0,54
V28/PCP	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.	80	K	1/78	5,77	0,019	0,56[a]	0,46
V29[b]	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	80	R/K	1/44	12,93	0,00081	0,81[a]	0,68

[a] Dieser Mittelwert entspricht der leichteren Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.37.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A2$ in SP_A unter Ausschluss von SP_{UZ}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F03[b]	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	64	R	1/37	11,03	0,002	4,06[a]	3,12
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	64	R	1/37	12,11	0,0013	4,06[a]	3,56
F12[b]	Ist die Suchmaschine erfolgreich?	64	R	1/37	11,49	0,0017	4,00[a]	3,44
F13[b]	Sind Sie mit der Suchmaschine zufrieden?	64	R	1/37	17,14	0,00019	4,12[a]	3,12
F16[b]	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	64	R	1/25	9,69	0,0045	5,41[a]	4,41
F18[b]	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	64	R	1/29	20,02	0,00011	5,44[a]	4,41
F24[b]	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	64	R	1/34	6,29	0,017	5,09[a]	4,38
F25[b]	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	64	R	1/30	27,69	$1{,}1 \cdot 10^{-05}$	5,25[a]	3,97
F26[b]	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	64	R	1/37	9,00	0,0048	4,41[a]	3,28
SK01-M[b]	Genauigkeit	64	K/R	1/62	11,54	0,0012	0,72[a]	0,54
SK04-M[b]	Suche	64	R/K	1/26	10,72	0,003	0,70[a]	0,56
SK08-M[b]	Suche	64	R	1/28	9,91	0,0039	0,74[a]	0,59
SK-E[c]	Ease of Use (EUCS)	64	R	1/35	9,56	0,0039	0,84[a]	0,68
SK-C[b]	Content (EUCS)	64	R	1/25	13,14	0,0013	0,75[a]	0,58
SK-E-88	EUCS-Skala-1988	64	K/R	1/62	8,91	0,0041	0,74[a]	0,60
SK-E-09[b]	EUCS-Skala-2009	64	K/R	1/62	8,25	0,0056	0,71[a]	0,57
SK-K[b]	Kriteriumsskala	64	R	1/32	12,76	0,0011	0,75[a]	0,57

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.

[b] Topiceffekt auch in SP_A vorhanden.

[c] Entspricht auch der Skala SK13-M.

Tab. C.38.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A2$ in SP_A unter Ausschluss von SP_{MV}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	96	R	1/55	6,13	0,016	3,83[a]	3,27
F03[b]	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	96	R	1/45	6,31	0,016	3,69[a]	3,23
F08[b]	Ist die Suchmaschine benutzerfreundlich?	96	R	1/57	18,48	$6{,}6 \cdot 10^{-05}$	4,23[a]	3,52
F12[b]	Ist die Suchmaschine erfolgreich?	96	R	1/52	6,22	0,016	3,83[a]	3,44
F13[b]	Sind Sie mit der Suchmaschine zufrieden?	96	R	1/57	7,66	0,0076	3,88[a]	3,15
F16[b]	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	96	R	1/39	9,11	0,0045	5,31[a]	4,56
F18[b]	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	96	R	1/38	10,18	0,0028	5,44[a]	4,56
F24[b]	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	96	R	1/47	17,14	0,00014	5,21[a]	4,46
F25[b]	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	96	R	1/51	6,67	0,013	5,00[a]	4,23

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.

Fortsetzung auf nächster Seite

[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.38 (Fortsetzung)

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F26[b]	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	96	R	1/57	13,32	0,00057	4,52[a]	3,42
SK01-M[b]	Genauigkeit	96	R/K	1/57	6,24	0,015	0,68[a]	0,55
SK04-M[b]	Suche	96	R	1/48	6,54	0,014	0,70[a]	0,59
SK-K[b]	Kriteriumsskala	96	R	1/47	13,08	0,00072	0,73[a]	0,57

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.39.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A2$ in SP_A unter Ausschluss von SP_{SB}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F13[b]	Sind Sie mit der Suchmaschine zufrieden?	32	R	1/17	17,57	0,00055	4,38[a]	3,00
F26[b]	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	32	R	1/17	31,02	$2{,}8 \cdot 10^{-05}$	4,88[a]	2,62
SK03-M	Benutzerfreundlichkeit	32	R/K	1/14	13,06	0,0028	0,79[a]	0,65
SK09-M[b]	Benutzerfreundlichkeit	32	R/K	1/16	17,49	0,00063	0,73[a]	0,55

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

Tab. C.40.: Signifikante Ergebnisse der Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit für $A2$ in SP_A unter Ausschluss von SP_{TD}. Berichtet wird jeweils die Stichprobe mit dem niedrigsten p-Wert. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde eine klassische (K) oder robuste (R) Varianzanalyse (V) durchgeführt.

				Gruppenunterschied			Mittelwert	
ID	**Beschreibung**	**n**	**V**	**df**	**F**	**p**	**Sonne**	**Wind**
F08[b]	Ist die Suchmaschine benutzerfreundlich?	80	R	1/46	10,02	0,0027	4,20[a]	3,58
F18[b]	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	80	R	1/36	13,66	0,00071	5,28[a]	4,55
F24[b]	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	80	R	1/45	7,34	0,0095	5,08[a]	4,35
F26[b]	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	80	R	1/45	10,26	0,0025	4,45[a]	3,33
SK08-M[b]	Suche	80	R/K	1/36	6,63	0,014	0,70[a]	0,59
SK-K[b]	Kriteriumsskala	80	R	1/39	7,87	0,0078	0,72[a]	0,60

[a] Dieser Mittelwert entspricht der zufriedenstellenderen Aufgabe.
[b] Topiceffekt auch in SP_A vorhanden.

C.5. Weitere Ergebnisse der Kovarianzanalyse

In diesem Anhang sind weitere Befunde der im zweiten Experiment durchgeführten Kovarianzanalysen aufgeführt. Neben einer Übersicht über Variablen, für die mit keiner der betrachteten

Kovariaten eine Änderung der im Kontext der Hauptanalyse berichteten Effekte zu beobachten ist (vgl. Abschn. C.5.1), sind in den folgenden beiden Abschnitten C.5.2 und C.5.3 die, in Abschnitt 6.4.6.2 aus Platzgründen nicht berichteten, Mittelwerte und Teststatistiken der signifikanten Kovarianzanalysen aufgeführt, für die eine Änderung der im Kontext der Hauptanalyse beschriebenen Effekte auftritt.

C.5.1. Variablen mit stabilen Befunden

Dieser Abschnitt stellt eine Übersicht derjenigen Variablen bereit, die im Rahmen der Kovarianzanalyse des zweiten Experiments über alle untersuchten möglichen zusätzlichen Einflussgrößen hinweg ein mit den Ergebnissen der Hauptauswertung übereinstimmendes Befundmuster aufweisen. Aufgrund der hohen Anzahl der Variablen sind die Ergebnisse getrennt nach Benutzerleistung und Benutzerzufriedenheit zusammengefasst.

Tab. C.41.: Übersicht über Benutzerleistungsvariablen, für die im Rahmen der Kovarianzanalyse für $A1$ in SP_A keine Effekte neu hinzukommen oder verschwinden. Für jedes Leistungsmaß ist jeweils die minimale und maximale Stichprobengröße über alle betrachteten Kovariaten hinweg angegeben.

ID	Beschreibung	n_{min}	n_{max}
M01	Anz. aufg. Dok.	96	108
M02	Anz. aufg. Dok. (erste 10 Dok.)	96	96
M03	Anz. aufg. Dok. (erste Suche)	96	108
M04	Anz. aufg. Dok. (letzte Suche)	96	108
M05	Anz. aufg. irrel. Dok.	96	96
M09	Anz. irrel. bew. Dok.	96	96
M10	Anz. rel. bew. Dok.	96	96
M12	Anz. rel. bew. Dok. (erste Suche)	96	96
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	96	96
M13	Anz. rel. bew. Dok. (letzte Suche)	96	96
M16	Anz. richtig rel. bew. Dok.	96	96
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	96	108
M18	Anz. richtig rel. bew. Dok. (erste Suche)	96	108
M19	Anz. richtig rel. bew. Dok. (letzte Suche)	96	96
B04	Durchschn. Bew. rel. Dok.	96	96
B05	Durchschn. Bew. rel. Dok. (erste Suche)	22	22
Z01	Durchschn. Betrachtungsz. aller Dok.	96	108
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	96	96
Z02-log	Durchschn. Betrachtungsz. falsch bew. Dok.	96	96
Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	96	96
Z05-log	Durchschn. Betrachtungsz. irrel. bew. Dok.	96	96
Z06	Durchschn. Betrachtungsz. irrel. Dok.	23	23
Z06-log	Durchschn. Betrachtungsz. irrel. Dok.	23	23
Z07	Durchschn. Betrachtungsz. rel. bew. Dok.	96	96
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	96	96
Z08-log	Durchschn. Betrachtungsz. rel. Dok.	96	96
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	96	96
Z10	Durchschn. Betrachtungsz. richtig irrel. bew. Dok.	20	20

Fortsetzung auf nächster Seite

Tab. C.41 (Fortsetzung)

ID	Beschreibung	n_{min}	n_{max}
Z10-log	Durchschn. Betrachtungsz. richtig irrel. bew. Dok.	20	20
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	96	96
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	96	96
V04	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	96	108
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	96
V06	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	96	96
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	96
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	96
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	96
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	96	108
V20	$\frac{\text{Anz. richtig rel. bew. Dok. (erste Suche)}}{\text{Anz. aufg. Dok. (erste Suche)}}$	22	22
V22	$\frac{\text{Anz. richtig rel. bew. Dok. (erste Suche)}}{\text{Anz. rel. Dok. im Korpus}}$	22	22
V23	$\frac{\text{Anz. richtig rel. bew. Dok. (erste Suche)}}{\text{Anz. zurückgeg. rel. Dok. (erste Suche)}}$	22	22
V24	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. aufg. Dok. (letzte Suche)}}$	22	22
V25	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. rel. bew. Dok. (letzte Suche)}}$	22	22
V26	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. rel. Dok. im Korpus}}$	22	22
V27	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. zurückgeg. rel. Dok. (letzte Suche)}}$	22	22
V32/BR	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	96	96
V33	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	96	96
S01	Anz. Suchen	96	96
S03	Letzte betr. Rankingpos.	96	96
S04	Suchdauer	96	108
S05	Zeit zum ersten richtig rel. bew. Dok.	96	108
S05-log	Zeit zum ersten richtig rel. bew. Dok.	96	108

Tab. C.42.: Übersicht über Benutzerzufriedenheitsvariablen, für die im Rahmen der Kovarianzanalyse für $A1$ in SP_A keine Effekte neu hinzukommen oder verschwinden. Für jedes Zufriedenheitsitem ist jeweils die minimale und maximale Stichprobengröße über alle betrachteten Kovariaten hinweg angegeben.

ID	Beschreibung	n_{min}	n_{max}
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	96	96
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	96	96
F08	Ist die Suchmaschine benutzerfreundlich?	96	96
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	96	96
F13	Sind Sie mit der Suchmaschine zufrieden?	96	96
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	96	108
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	96	96
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	96	96
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	96	96
F22	Ich bin mit den Suchergebnissen zufrieden.	96	96
F23	Ich bin mit meiner Suchleistung zufrieden.	96	96
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	96	96
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	96	96
SK01-M	Genauigkeit	96	96
SK03-M	Benutzerfreundlichkeit	96	96
SK07-M[a]	Benutzerfreundlichkeit	96	108
SK08-M	Suche	96	96
SK09-M	Benutzerfreundlichkeit	96	96
SK-E[b]	Ease of Use (EUCS)	96	96
SK-A	Accuracy (EUCS)	96	96
SK-T	Timeliness (EUCS)	96	108
SK-K	Kriteriumsskala	96	96
SK-E-88	EUCS-Skala-1988	96	96
SK-E-09	EUCS-Skala-2009	96	96

[a] Entspricht auch der Skala SK18-M.
[b] Entspricht auch der Skala SK13-M.

C.5.2. Gruppenmittelwerte der Kovarianzanalyse

Dieser Abschnitt beinhaltet die innerhalb der Arbeit aus Platzgründen nicht berichteten Mittelwerte der signifikanten Kovarianzanalysen für die eine Änderung der im Kontext der Hauptanalyse beschriebenen Effekte zu beobachten ist. Wie in Abschnitt 6.4.6.2 erläutert, erfolgt die Überprüfung der Stabilität der Befunde im zweiten Experiment aufgrund zu geringer Stichprobengrößen im Fall von Aufgabe 2 nur für die erste von den Testpersonen bearbeitete Aufgabe. Die entsprechenden Analyseergebnisse für Aufgabe 1 können im Folgenden getrennt nach Benutzerleistung und Benutzerzufriedenheit nachvollzogen werden.

Tab. C.43.: Signifikante Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses personenbezogener Störfaktoren auf die Benutzerleistung für $A1$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

			System		Erwartung		Interaktion			
ID	**Beschreibung**	**Kov**	**S_G**	**S_S**	**E_H**	**E_N**	**$I_{G,H}$**	**$I_{G,N}$**	**$I_{S,H}$**	**$I_{S,N}$**
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	K04	**0,71**[b]	**0,78**	0,74	0,75	0,70	0,71	0,77	0,78
		K05	**0,70**[b]	**0,80**	0,73	0,77	0,69	0,72	0,78	0,81
		K06	**0,66**[b]	**0,73**	0,69	0,70	0,65	0,67	0,73	0,74
		K08	**0,76**[b]	**0,84**	0,80	0,80	0,76	0,76	0,84	0,84
		K10	**0,75**[b]	**0,83**	0,79	0,79	0,75	0,75	0,83	0,83
		K12	**0,71**[b]	**0,80**	0,76	0,76	0,71	0,71	0,80	0,80
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	K08	0,47	0,80	0,61	0,66	**0,50**[a]	**0,43**	**0,72**	**0,89**
V18	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. aufg. Dok. (erste 10 Dok.)	K06	**0,56**[a]	**0,42**	0,50	0,49	0,54	0,58	0,46	0,39
V19	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)	K05	**1,00**[b]	**0,92**	0,96	0,96	1,00	1,00	0,92	0,92
V21	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. bew. Dok. (erste Suche)	K04	0,96	0,88	0,89	0,95	**0,97**[a]	**0,96**	**0,81**	**0,95**
		K05	0,89	0,81	0,81	0,89	**0,89**[a]	**0,89**	**0,74**	**0,89**
		K06	0,89	0,81	**0,81**[a]	**0,89**	**0,89**[a]	**0,89**	**0,74**	**0,89**
		K10	0,89	0,81	0,81	0,89	**0,89**[a]	**0,89**	**0,74**	**0,89**
V28/PCP	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.	K08	**0,56**[b]	**0,51**	0,52	0,55	0,55	0,58	0,50	0,52
		K12	**0,53**[b]	**0,48**	0,49	0,51	0,51	0,55	0,47	0,48
V29	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	K12	**0,73**[a]	**0,86**	0,77	0,82	0,71	0,75	0,82	0,89

[a] Effekt kommt in der Kovarianzanalyse hinzu.
[b] Effekt entfällt in der Kovarianzanalyse.

Tab. C.44.: Signifikante Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses personenbezogener Störfaktoren auf die Benutzerzufriedenheit für $A1$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

			System		Erwartung		Interaktion			
ID	**Beschreibung**	**Kov**	**S_G**	**S_S**	**E_H**	**E_N**	**$I_{G,H}$**	**$I_{G,N}$**	**$I_{S,H}$**	**$I_{S,N}$**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	K04	4,00	3,50	**4,00**[a]	**3,50**	4,00	4,00	4,00	3,00
		K05	4,00	3,50	**4,00**[a]	**3,50**	4,00	4,00	4,00	3,00
		K06	4,00	3,50	**4,00**[a]	**3,50**	4,00	4,00	4,00	3,00
		K09	4,00	3,50	**4,00**[a]	**3,50**	4,00	4,00	4,00	3,00
		K10	4,00	3,50	**4,00**[a]	**3,50**	4,00	4,00	4,00	3,00

[a] Effekt kommt in der Kovarianzanalyse hinzu.
[b] Effekt entfällt in der Kovarianzanalyse.

Fortsetzung auf nächster Seite

Tab. C.44 (Fortsetzung)

			System		Erwartung		Interaktion			
ID	**Beschreibung**	**Kov**	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$	$\mathbf{I_{G,H}}$	$\mathbf{I_{G,N}}$	$\mathbf{I_{S,H}}$	$\mathbf{I_{S,N}}$
F09	Ist die Suchmaschine einfach zu bedienen?	K01	5,00	5,00	**5,00**[b]	**5,00**	5,00	5,00	5,00	5,00
		K04	5,00	5,00	**5,00**[b]	**5,00**	5,00	5,00	5,00	5,00
		K05	5,00	5,00	**5,00**[b]	**5,00**	5,00	5,00	5,00	5,00
		K06	5,00	5,00	**5,00**[b]	**5,00**	5,00	5,00	5,00	5,00
		K08	5,00	5,00	**5,00**[b]	**5,00**	5,00	5,00	5,00	5,00
		K09	5,00	5,00	**5,00**[b]	**5,00**	5,00	5,00	5,00	5,00
		K12	5,00	5,00	**5,00**[b]	**5,00**	5,00	5,00	5,00	5,00
F11	Liefert die Suchmaschine aktuelle Information?	K04	**3,50**[a]	**4,00**	4,00	3,50	4,00	3,00	4,00	4,00
		K04	3,50	4,00	4,00	3,50	**4,00**[a]	**3,00**	**4,00**	**4,00**
		K05	**3,50**[a]	**4,00**	4,00	3,50	4,00	3,00	4,00	4,00
		K05	3,50	4,00	4,00	3,50	**4,00**[a]	**3,00**	**4,00**	**4,00**
		K06	**3,50**[a]	**4,00**	4,00	3,50	4,00	3,00	4,00	4,00
		K06	3,50	4,00	4,00	3,50	**4,00**[a]	**3,00**	**4,00**	**4,00**
		K08	**3,50**[a]	**4,00**	4,00	3,50	4,00	3,00	4,00	4,00
		K08	3,50	4,00	4,00	3,50	**4,00**[a]	**3,00**	**4,00**	**4,00**
		K12	**3,50**[a]	**4,00**	4,00	3,50	4,00	3,00	4,00	4,00
		K12	3,50	4,00	4,00	3,50	**4,00**[a]	**3,00**	**4,00**	**4,00**
F12	Ist die Suchmaschine erfolgreich?	K04	**4,00**[a]	**4,50**	4,50	4,00	**4,00**[a]	**4,00**	**5,00**	**4,00**
		K05	**4,00**[a]	**4,50**	4,50	4,00	**4,00**[a]	**4,00**	**5,00**	**4,00**
		K06	**4,00**[a]	**4,50**	4,50	4,00	**4,00**[a]	**4,00**	**5,00**	**4,00**
		K08	**4,00**[a]	**4,50**	4,50	4,00	**4,00**[a]	**4,00**	**5,00**	**4,00**
		K10	**4,00**[a]	**4,50**	4,50	4,00	**4,00**[a]	**4,00**	**5,00**	**4,00**
F14	Es war einfach, die Aufgabe zu bearbeiten.	K11	4,67	4,58	**4,92**[a]	**4,33**	4,83	4,50	5,00	4,17
SK02-M	Inhalt	K05	0,60	0,63	**0,66**[b]	**0,57**	0,64	0,56	0,68	0,58
		K01	24,30	24,35	**24,36**[b]	**24,29**	24,34	24,27	24,39	24,30
SK04-M	Suche	K02	0,69	0,75	**0,75**[b]	**0,69**	0,72	0,67	0,78	0,72
SK05-F	Aufgabe	K02	0,13	0,04	**0,27**[a]	**−0,10**	0,26	$9{\cdot}10^{-4}$	0,28	−0,20
		K06	−0,47	−0,52	**−0,30**[a]	**−0,69**	−0,32	−0,62	−0,27	−0,76
		K08	0,04	−0,05	**0,19**[a]	**−0,20**	0,19	−0,11	0,19	−0,29
		K12	0,07	0,04	**0,27**[a]	**−0,16**	0,24	−0,10	0,30	−0,21
SK05-M	Aufgabe	K06	0,65	0,63	**0,69**[a]	**0,58**	0,69	0,61	0,69	0,56
		K08	0,75	0,75	**0,79**[a]	**0,71**	0,79	0,71	0,79	0,71
		K12	0,79	0,79	**0,83**[a]	**0,75**	0,83	0,75	0,83	0,75

[a] Effekt kommt in der Kovarianzanalyse hinzu.

[b] Effekt entfällt in der Kovarianzanalyse.

Fortsetzung auf nächster Seite

Tab. C.44 (Fortsetzung)

			System		Erwartung		Interaktion			
ID	**Beschreibung**	**Kov**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
SK06-M	Eigenleistung	K01	0,44	0,50	**0,53**[b]	**0,42**	0,50	0,39	0,56	0,44
		K02	0,44	0,50	**0,53**[b]	**0,42**	0,50	0,39	0,55	0,44
		K04	0,38	0,44	**0,44**[b]	**0,37**	0,41	0,34	0,48	0,40
		K05	0,44	0,51	**0,52**[b]	**0,44**	0,48	0,41	0,55	0,47
		K06	0,47	0,44	**0,50**[b]	**0,42**	0,50	0,44	0,50	0,39
		K07	0,42	0,42	**0,47**[b]	**0,36**	0,44	0,39	0,50	0,33
		K08	0,39	0,44	**0,46**[b]	**0,38**	0,43	0,36	0,48	0,40
		K09	0,38	0,44	**0,45**[b]	**0,37**	0,42	0,34	0,48	0,40
		K10	0,44	0,47	**0,50**[b]	**0,42**	0,47	0,42	0,53	0,42
SK10-F	Aufgabe	K06	−0,56	−0,61	**−0,30**[a]	**−0,87**	−0,31	−0,81	−0,29	−0,94
SK11-M	Eigenleistung	K07	0,54	0,46	**0,58**[a]	**0,42**	0,58	0,50	0,58	0,33
SK-C	Content (EUCS)	K01	17,35	17,39	**17,41**[b]	**17,33**	17,38	17,32	17,45	17,34
		K02	0,66	0,69	**0,72**[b]	**0,62**	0,69	0,62	0,75	0,63
		K05	0,62	0,64	**0,67**[b]	**0,59**	0,66	0,59	0,68	0,60
		K10	0,64	0,68	**0,71**[b]	**0,62**	0,68	0,61	0,74	0,62

[a] Effekt kommt in der Kovarianzanalyse hinzu.
[b] Effekt entfällt in der Kovarianzanalyse.

C.5.3. Teststatistiken der Kovarianzanalyse

Die in diesem Abschnitt dargestellten Tabellen enthalten die Teststatistiken zu im vorangegangenen Abschnitt aufgeführten Gruppenmittelwerten. Die Darstellung erfolgt erneut getrennt nach Benutzerleistung und Benutzerzufriedenheit.

Tab. C.45.: Teststatistik der Kovarianzanalyse zur Untersuchung des Einflusses personenbezogener Störfaktoren auf die Benutzerleistung für $A1$ in SP_A. Fett hervorgehoben sind Effekte bei denen sich die Ergebnisse der Kovarianzanalyse von denen der Kovarianzanalyse unterscheiden. Ist ein Effekt in der Kovarianzanalyse in mindestens vier der fünf Zufallsstichproben signifikant, wird der minimale ansonsten der maximale p-Wert berichtet. Fett gedruckte p-Werte kleiner gleich als $0{,}05$ weisen somit auf in der Kovarianzanalyse neu hinzukommende Effekte hin, während fett gedruckte p-Werte größer als $0{,}05$ entfallende Effekte kennzeichnen.

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	K04	88	0,26	0,61	2,09	**0,15**[a]	0,07	0,79	$-1{\cdot}10^{-4}$	1,00
		K05	88	1,87	0,17	3,25	**0,08**[a]	0,54	0,47	0,003	0,96
		K06	88	1,65	0,20	2,48	**0,12**[a]	0,04	0,85	0,001	0,97
		K08	88	0,44	0,51	2,68	**0,11**[a]	0,01	0,91	0,009	0,92
		K10	88	$2{\cdot}10^{-5}$	1,00	2,24	**0,14**[a]	$-8{\cdot}10^{-8}$	1,00	$-1{\cdot}10^{-7}$	1,00
		K12	88	1,34	0,25	2,42	**0,12**[a]	$2{\cdot}10^{-6}$	1,00	$-1{\cdot}10^{-5}$	1,00
V17	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	K08	96	3,78	0,05	24,48	$3{\cdot}10^{-6}$	0,97	0,33	4,53	**0,04**[b,d]

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

Fortsetzung auf nächster Seite

Tab. C.45 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
V18	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. aufg. Dok. (erste 10 Dok.)	K06	88	0,38	0,54	8,63	**0,004**[b,d]	0,06	0,80	1,34	0,25
V19	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)	K05	88	0,001	0,97	1,70	**0,20**[a]	$-1{\cdot}10^{-5}$	1,00	$-1{\cdot}10^{-5}$	1,00
V21	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. bew. Dok. (erste Suche)	K04	88	2,91	0,09	13,55	$4{\cdot}10^{-4}$	6,11	0,02	10,45	**0,002**[b,d]
		K05	88	0,002	0,96	20,55	$2{\cdot}10^{-5}$ [c]	7,64	0,007	13,34	**$5{\cdot}10^{-4}$**[b,d]
		K06	88	$6{\cdot}10^{-4}$	0,98	20,39	$2{\cdot}10^{-5}$ [c]	10,33	**0,002**[b,d]	14,11	**$3{\cdot}10^{-4}$**[b,d]
		K10	88	0,002	0,97	20,55	$2{\cdot}10^{-5}$ [c]	9,20	0,003	13,92	**$3{\cdot}10^{-4}$**[b,d]
V28/PCP	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.	K08	96	0,42	0,52	1,54	**0,22**[a]	0,34	0,56	0,04	0,85
		K12	96	0,94	0,34	1,96	**0,16**[a]	0,22	0,64	0,10	0,75
V29	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	K12	96	$5{\cdot}10^{-5}$	0,99	5,75	**0,02**[b,d]	0,78	0,38	0,12	0,73

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

Tab. C.46.: Teststatistik der Kovarianzanalyse zur Untersuchung des Einflusses personenbezogener Störfaktoren auf die Benutzerzufriedenheit für $A1$ in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Ist ein Effekt in der Kovarianzanalyse in mindestens vier der fünf Zufallsstichproben signifikant, wird der minimale ansonsten der maximale p-Wert berichtet. Fett gedruckte p-Werte kleiner gleich als $0{,}05$ weisen somit auf in der Kovarianzanalyse neu hinzukommende Effekte hin, während fett gedruckte p-Werte größer als $0{,}05$ entfallende Effekte kennzeichnen.

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	K04	108	0,06	0,81	7,30	0,008	43,73	**$2{\cdot}10^{-9}$**[b,c]	7,30	0,008
		K05	108	0,06	0,81	7,30	0,008	38,66	**$1{\cdot}10^{-8}$**[b,c]	7,30	0,008
		K06	108	0,06	0,81	7,30	0,008	43,76	**$2{\cdot}10^{-9}$**[b,c]	7,30	0,008
		K09	108	0,06	0,81	7,30	0,008	43,73	**$2{\cdot}10^{-9}$**[b,c]	7,30	0,008
		K10	108	0,06	0,81	7,30	0,008	38,90	**$1{\cdot}10^{-8}$**[b,c]	7,30	0,008
F09	Ist die Suchmaschine einfach zu bedienen?	K01	96	$5{\cdot}10^{-5}$	0,99	$-8{\cdot}10^{-6}$	1,00	$-3{\cdot}10^{-6}$	**1,00**[a]	$-3{\cdot}10^{-6}$	1,00
		K04	96	$4{\cdot}10^{-5}$	0,99	$-2{\cdot}10^{-5}$	1,00	$-9{\cdot}10^{-6}$	**1,00**[a]	$4{\cdot}10^{-6}$	1,00
		K05	96	$-2{\cdot}10^{-6}$	1,00	$-3{\cdot}10^{-6}$	1,00	$-3{\cdot}10^{-6}$	**1,00**[a]	$1{\cdot}10^{-5}$	1,00
		K06	96	$5{\cdot}10^{-5}$	0,99	$-2{\cdot}10^{-6}$	1,00	$-5{\cdot}10^{-6}$	**1,00**[a]	$-2{\cdot}10^{-6}$	1,00
		K08	96	$-5{\cdot}10^{-6}$	1,00	$-2{\cdot}10^{-5}$	1,00	$-1{\cdot}10^{-5}$	**1,00**[a]	$-2{\cdot}10^{-5}$	1,00
		K09	96	$5{\cdot}10^{-5}$	0,99	$-1{\cdot}10^{-5}$	1,00	$-2{\cdot}10^{-5}$	**1,00**[a]	$-3{\cdot}10^{-6}$	1,00
		K12	96	$6{\cdot}10^{-5}$	0,99	$-3{\cdot}10^{-6}$	1,00	$-3{\cdot}10^{-6}$	**1,00**[a]	$-2{\cdot}10^{-6}$	1,00

Fortsetzung auf nächster Seite

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

Tab. C.46 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
F11	Liefert die Suchmaschine aktuelle Information?	K04	96	$-2{\cdot}10^{-5}$	1,00	55,25	$\mathbf{6{\cdot}10^{-11}}$ b,d	87,90	$5{\cdot}10^{-15}$ c	32,23	$2{\cdot}10^{-7}$
		K04	96	$-4{\cdot}10^{-6}$	1,00	46,35	$1{\cdot}10^{-9}$	74,91	$2{\cdot}10^{-13}$	38,28	$\mathbf{2{\cdot}10^{-8}}$ b,d
		K05	96	$-3{\cdot}10^{-5}$	1,00	55,25	$\mathbf{6{\cdot}10^{-11}}$ b,d	87,90	$5{\cdot}10^{-15}$ c	32,23	$2{\cdot}10^{-7}$
		K05	96	$-1{\cdot}10^{-5}$	1,00	47,41	$7{\cdot}10^{-10}$	74,91	$2{\cdot}10^{-13}$	38,28	$\mathbf{2{\cdot}10^{-8}}$ b,d
		K06	96	$2{\cdot}10^{-5}$	1,00	55,25	$\mathbf{6{\cdot}10^{-11}}$ b,d	87,90	$5{\cdot}10^{-15}$ c	32,23	$2{\cdot}10^{-7}$
		K06	96	$1{\cdot}10^{-6}$	1,00	46,35	$1{\cdot}10^{-9}$	74,91	$2{\cdot}10^{-13}$	38,29	$\mathbf{2{\cdot}10^{-8}}$ b,d
		K08	96	$-6{\cdot}10^{-6}$	1,00	55,25	$\mathbf{6{\cdot}10^{-11}}$ b,d	91,01	$2{\cdot}10^{-15}$ c	32,23	$2{\cdot}10^{-7}$
		K08	96	$-9{\cdot}10^{-6}$	1,00	46,35	$1{\cdot}10^{-9}$	74,91	$2{\cdot}10^{-13}$	38,29	$\mathbf{2{\cdot}10^{-8}}$ b,d
		K12	96	$3{\cdot}10^{-5}$	1,00	55,25	$\mathbf{6{\cdot}10^{-11}}$ b,d	87,90	$5{\cdot}10^{-15}$ c	32,23	$2{\cdot}10^{-7}$
		K12	96	$-7{\cdot}10^{-6}$	1,00	46,35	$1{\cdot}10^{-9}$	74,91	$2{\cdot}10^{-13}$	38,28	$\mathbf{2{\cdot}10^{-8}}$ b,d
F12	Ist die Suchmaschine erfolgreich?	K04	96	$-3{\cdot}10^{-4}$	1,00	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,c	18,89	$4{\cdot}10^{-5}$ c	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,c
		K05	96	$5{\cdot}10^{-5}$	0,99	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,d	18,89	$4{\cdot}10^{-5}$ c	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,d
		K06	96	$5{\cdot}10^{-5}$	0,99	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,c	18,89	$4{\cdot}10^{-5}$ c	18,92	$\mathbf{4{\cdot}10^{-5}}$ b,c
		K08	96	$6{\cdot}10^{-5}$	0,99	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,d	18,89	$4{\cdot}10^{-5}$ c	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,d
		K10	96	$1{\cdot}10^{-5}$	1,00	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,c	18,89	$4{\cdot}10^{-5}$ c	18,89	$\mathbf{4{\cdot}10^{-5}}$ b,d
F14	Es war einfach, die Aufgabe zu bearbeiten.	K11	96	14,19	$\mathbf{3{\cdot}10^{-4}}$ c	0,63	0,43	11,81	$\mathbf{9{\cdot}10^{-4}}$ b,c	2,47	0,12
SK02-M	Inhalt	K01	96	2,13	0,15	1,11	0,29	3,12	**0,08** a	0,04	0,84
		K05	96	1,03	0,31	0,59	0,44	3,27	**0,07** a	0,07	0,80
SK04-M	Suche	K02	96	0,64	0,43	1,08	0,30	2,14	**0,15** a	$2{\cdot}10^{-6}$	1,00
SK05-F	Aufgabe	K02	108	13,59	$\mathbf{4{\cdot}10^{-4}}$ c	0,44	0,51	7,52	**0,007** b,c	0,69	0,41
		K06	108	9,52	**0,003** c	0,08	0,78	6,54	**0,01** b,c	0,25	0,62
		K08	108	3,35	0,07	0,31	0,58	5,91	**0,02** b,d	0,30	0,59
		K12	108	1,82	0,18	0,02	0,89	6,27	**0,01** b,d	0,31	0,58
SK05-M	Aufgabe	K06	96	10,18	**0,002** d	1,48	0,23	13,40	$\mathbf{4{\cdot}10^{-4}}$ b,c	0,87	0,35
		K08	96	3,12	0,08	$3{\cdot}10^{-5}$	1,00	13,97	$\mathbf{3{\cdot}10^{-4}}$ b,d	$-2{\cdot}10^{-5}$	1,00
		K12	96	$1{\cdot}10^{-6}$	1,00	0,004	0,95	9,63	**0,003** b,d	$2{\cdot}10^{-4}$	0,99
SK06-M	Eigenleistung	K01	96	0,008	0,93	0,93	0,34	1,44	**0,23** a	$9{\cdot}10^{-5}$	0,99
		K02	96	2,28	0,13	0,92	0,34	1,71	**0,19** a	0,02	0,88
		K04	96	1,45	0,23	1,59	0,21	2,39	**0,13** a	0,008	0,93
		K05	96	0,14	0,71	1,60	0,21	2,00	**0,16** a	0,02	0,89
		K06	96	$-6{\cdot}10^{-6}$	1,00	0,15	0,70	2,73	**0,10** a	0,15	0,70
		K07	96	4,05	**0,05** d	$-4{\cdot}10^{-6}$	1,00	3,13	**0,08** a	0,32	0,57
		K08	96	7,99	**0,006** d	0,58	0,45	2,04	**0,16** a	0,006	0,94
		K09	96	1,06	0,31	1,19	0,28	2,31	**0,13** a	$-2{\cdot}10^{-4}$	1,00
		K10	96	0,004	0,95	0,04	0,85	1,94	**0,17** a	0,18	0,67
SK10-F	Aufgabe	K06	96	9,50	**0,003** c	0,03	0,86	9,99	**0,002** b,d	0,18	0,67
SK11-M	Eigenleistung	K07	96	7,54	**0,007** c	1,72	0,19	9,96	**0,002** b,d	2,06	0,15
SK-C	Content (EUCS)	K01	96	0,67	0,42	1,03	0,31	3,00	**0,09** a	0,21	0,65
		K02	96	$-3{\cdot}10^{-8}$	1,00	0,98	0,32	3,73	**0,06** a	0,42	0,52
		K05	96	2,09	0,15	0,14	0,71	3,18	**0,08** a	0,10	0,75
		K10	96	0,16	0,69	0,96	0,33	3,61	**0,06** a	0,24	0,62

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

D. Verwendete Materialien für Experiment 3

Dieser Anhang enthält die Materialien, die im dritten Experiment dieser Arbeit verwendet werden. Dies sind neben den verschiedenen Instruktionstexten zur Manipulation der Erwartungshaltung auch die im dritten Experiment aufgrund der 4-stufigen Relevanzskala zusätzlich verwendeten Leistungsmaße zur Beurteilung des Sucherfolgs der Testpersonen, eine Übersicht über die eingesetzten Zufriedenheitsskalen sowie die im Rahmen der Kovarianzanalyse zusätzlich berücksichtigten Benutzerleistungskovariaten.

D.1. Testinstruktion

In diesem Abschnitt sind die Instruktionstexte angegeben, die die Probanden auf die Aufgaben des Experiments vorbereiten. Als Erwartungsmanipulation erhalten die Teilnehmer im Rahmen des dritten Experiments alle denselben Instruktionstext. Zum besseren Vergleich der drei Experimente gliedert sich der vorliegende Abschnitt in die drei Bereiche Einführung, Erwartungsmanipulation und Aufgabenbeschreibung.

D.1.1. Einführung des dritten Experiments

Herzlich willkommen und vielen Dank für Ihre Bereitschaft zur Teilnahme an unserer Untersuchung. In dem heutigen Benutzertest sollen zwei Suchmaschinen miteinander verglichen werden. Insgesamt wird der Test ca. 45 Minuten in Anspruch nehmen. Bitte beachten Sie, in dieser Untersuchung geht es nicht um eine Beurteilung Ihrer Person, sondern lediglich um Ihre persönliche Bewertung einer der beiden Suchmaschinen. Alle Ihre Antworten werden selbstverständlich anonym erhoben und ausschließlich zu wissenschaftlichen Zwecken ausgewertet.

D.1.2. Erwartungsmanipulation des dritten Experiments

Aktuelle Studien zur Suchmaschinennutzung haben gezeigt, dass die Suchmaschinenleistung keinen großen Einfluss auf den Sucherfolg der Benutzer hat. Vielmehr erreichen Benutzer auch mit unterschiedlich guten Suchmaschinen vergleichbare Ergebnisse. Wir möchten dem heute nachgehen, indem wir zwei unterschiedlich gut bewertete Suchsysteme miteinander vergleichen. Um den Aufwand für Sie so gering wie möglich zu halten, haben wir uns dafür entschieden, die Suchmaschinen von zwei unabhängigen Gruppen evaluieren zu lassen. Die Gruppenzuteilung erfolgte per Losverfahren und Sie wurden dem besseren/schlechtere der beiden Systeme zugeteilt. Im Folgenden werden wir Ihnen den genauen Ablauf des Tests erläutern.

D.1.3. Aufgabenbeschreibung des dritten Experiments

Ihnen werden sogleich nacheinander drei Suchthemen präsentiert und Ihre Aufgabe wird es sein, die Qualität der gefundenen Dokumente zu bewerten. Die maximale Bearbeitungszeit beträgt zehn Minuten pro Thema. Sollten Sie sich bereits vor Ablauf dieser Zeitspanne ausreichend informiert fühlen, können Sie jederzeit mit der nächsten Aufgabe fortfahren. Nach Eingabe einer Suchanfrage wird eine Ergebnisliste mit den gefundenen Dokumenten ausgegeben. Jede Ergebnisliste besteht aus mehreren Ergebnisseiten zwischen denen Sie über die Seitenauswahl vor und zurück wechseln können. Des Weiteren können Sie auch die Aufgabenbeschreibung jederzeit noch einmal aufrufen. Selektieren Sie bitte nur Links, die Ihrer Ansicht nach auf relevante Dokumente verweisen. Sobald Sie einen Link anklicken, wird das jeweilige Webdokument in einem neuen Tab geöffnet. Nachdem Sie das Dokument betrachtet haben, kehren Sie zum Tab mit der Suchmaschine zurück. Auf der dortigen Skala haben Sie die Möglichkeit Ihr Relevanzurteil für

das betrachtete Webdokument abzugeben, bevor Sie über den Speicher-Button wieder zurück zur Ergebnisliste gelangen. Bevor es nun tatsächlich losgeht, noch ein kurzer Hinweis. Alle Teilnehmer haben die Chance einen von drei Geldpreisen im Wert von 20, 30 oder 50€ zu gewinnen. Die Gewinner werden aus den jeweils zehn besten Teilnehmern beider Suchmaschinen ausgelost und schriftlich benachrichtigt. Und nun viel Spaß bei der ersten Aufgabe!

D.2. Skalen zur Beurteilung der Benutzerzufriedenheit

Dieser Abschnitt bietet eine Übersicht über die im dritten Experiment ermittelten Zufriedenheitsskalen. Die empirische Herleitung dieser Skalen ist ausführlich in Abschnitt 7.4.4.2 beschrieben. Neben den im Rahmen der Faktorenanalyse ermittelten Skalen gibt Tabelle D.1 auch Aufschluss über die im Zuge der Auswertung gebildeten Gesamtskalen (z.B. SK-E-88) und über Skalen, die zum Vergleich mit Experiment 2 aus dem zweiten Experiment übernommen werden (z.B. SK01-M).

Tab. D.1.: Übersicht über verwendete Benutzerzufriedenheitsskalen. Die Buchstaben M bzw. F am Ende der Skalenbezeichnung geben an, ob zur Berechnung der Mittelwert der Items (M) oder der entsprechende Faktorwert (F) zugrunde gelegt wird. Aus Experiment 2 übernommene Skalen sind mit einer Fußnote gekennzeichnet.

Skala	**Beschreibung**	**Items**
SK01-M[a]	Genauigkeit	F02, F03, F04
SK02-M[a]	Inhalt	F01, F05, F06
SK03-M[a]	Benutzerfreundlichkeit	F07, F08
SK04-M[a]	Suche	F16, F18, F19
SK07-M[a,c]	Benutzerfreundlichkeit	F17, F24
SK08-M[a]	Suche	F01, F02, F03, F04, F05, F16, F17, F18, F19
SK09-M[a]	Benutzerfreundlichkeit	F07, F08, F24
SK11-M[a,b]	Eigenleistung	F20, F23
SK12-M/SK12-F	Suchergebnis	F01, F02, F03, F04, F05, F06, F07
SK13-M[d]/SK13-F	Benutzerfreundlichkeit	F08, F09
SK14-M/SK14-F	Suche	F17, F18, F19
SK15-M[b]/SK15-F	Eigenleistung	F20, F23
SK16-M/SK16-F	Aufgabe	F14, F16
SK17-M/SK17-F	Suche	F01, F03, F04, F05, F07, F18, F19
SK18-M[c]/SK18-F	Benutzerfreundlichkeit	F17, F24
SK19-M[b]/SK19-F	Eigenleistung	F20, F23
SK-A[a]	Accuracy (EUCS)	F05, F06
SK-C[a]	Content (EUCS)	F01, F02, F03, F04
SK-E[a,d]	Ease of Use (EUCS)	F08, F09
SK-T[a]	Timeliness (EUCS)	F10, F11
SK-K[a]	Kriteriumsskala	F12, F13
SK-E-88[a]	EUCS-Skala-1988	F01, F02, F03, F04, F05, F06, F08, F09, F10, F11
SK-E-09[a]	EUCS-Skala-2009	F01, F02, F03, F04, F05, F06, F07, F08
SK-E-13	EUCS-Skala-2013	F01, F02, F03, F04, F05, F06, F07, F08, F09
SK-Z-13	Zusatzskala-2013	F14, F16, F17, F18, F19, F20, F23, F24
SK-G-13	Gesamtskala-2013	F01, F03, F04, F05, F07, F14, F17, F18, F19, F20, F23, F24

[a] Skala aus Experiment 2 übernommen.
[b] Die Mittelwertsskalen SK11-M, SK15-M und SK19-M sind identisch.
[c] Die Mittelwertsskalen SK07-M und SK18-M sind identisch.
[d] Die Mittelwertsskalen SK13-M und SK-E sind identisch.

D.3. Maße zur Beurteilung der Benutzerleistung

Dieser Abschnitt bietet eine Übersicht über die im dritten Experiment aufgrund der 4-stufigen Relevanzskala hinzukommenden Benutzerleistungsmaße. Entsprechend ihrer Einführung im Rahmen der Operationalisierung der abhängigen Variablen (vgl. Abschn. 7.3.2) sind die in den Tabellen D.2 bis D.6 beschriebenen Benutzerleistungsmaße in folgende fünf Variablengruppen untergliedert: Dokumentenmengen, durchschnittliche Bewertungen von Dokumentenmengen, durchschnittliche Betrachtungszeiten unterschiedlicher Dokumentenmengen, Verhältnisse von Dokumentenmengen und sonstige Leistungsmaße.

Tab. D.2.: Dokumentenmengen bei 4-stufiger Relevanzskala (Variablengrp. 1). Bei der Zählung der Dokumentenmengen wird jedes Dokument nur einmal berücksichtigt. Als Dokumentenbewertung geht jeweils die letzte Bewertung eines Dokuments über alle Suchanfragen hinweg in die Betrachtung ein.

ID	Beschreibung	Kurzform
M20	Anzahl aufgerufener eher irrelevanter Dokumente	Anz. aufg. eher irrel. Dok.
M21	Anzahl aufgerufener eher relevanter Dokumente	Anz. aufg. eher rel. Dok.
M22	Anzahl aufgerufener irrelevanter Dokumente	Anz. aufg. irrel. Dok.
M23	Anzahl aufgerufener relevanter Dokumente	Anz. aufg. rel. Dok.
M24	Anzahl eher irrelevant bewerteter Dokumente	Anz. eher irrel. bew. Dok.
M25	Anzahl eher relevant bewerteter Dokumente	Anz. eher rel. bew. Dok.
M26	Anzahl falsch eher irrelevant bewerteter Dokumente	Anz. falsch eher irrel. bew. Dok.
M27	Anzahl falsch eher irrelevant bewerteter eher relevanter Dokumente	Anz. falsch eher irrel. bew. eher rel. Dok.
M28	Anzahl falsch eher irrelevant bewerteter irrelevanter Dokumente	Anz. falsch eher irrel. bew. irrel. Dok.
M29	Anzahl falsch eher irrelevant bewerteter relevanter Dokumente	Anz. falsch eher irrel. bew. rel. Dok.
M30	Anzahl falsch eher relevant bewerteter Dokumente	Anz. falsch eher rel. bew. Dok.
M31	Anzahl falsch eher relevant bewerteter eher irrelevanter Dokumente	Anz. falsch eher rel. bew. eher irrel. Dok.
M32	Anzahl falsch eher relevant bewerteter irrelevanter Dokumente	Anz. falsch eher rel. bew. irrel. Dok.
M33	Anzahl falsch eher relevant bewerteter relevanter Dokumente	Anz. falsch eher rel. bew. rel. Dok.
M34	Anzahl falsch irrelevant bewerteter Dokumente	Anz. falsch irrel. bew. Dok.
M35	Anzahl falsch relevant bewerteter Dokumente	Anz. falsch rel. bew. Dok.
M36	Anzahl irrelevant bewerteter Dokumente	Anz. irrel. bew. Dok.
M37	Anzahl relevant bewerteter Dokumente	Anz. rel. bew. Dok.
M38	Anzahl relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente	Anz. rel. bew. Dok. (erste 10 Dok.)
M39	Anzahl relevant bewerteter Dokumente der ersten durchgeführten Suche	Anz. rel. bew. Dok. (erste Suche)
M40	Anzahl relevant bewerteter Dokumente der letzten durchgeführten Suche	Anz. rel. bew. Dok. (letzte Suche)
M41	Anzahl richtig bewerteter Dokumente	Anz. richtig bew. Dok.
M42	Anzahl richtig eher irrelevant bewerteter Dokumente	Anz. richtig eher irrel. bew. Dok.
M43	Anzahl richtig eher relevant bewerteter Dokumente	Anz. richtig eher rel. bew. Dok.
M44	Anzahl richtig irrelevant bewerteter Dokumente	Anz. richtig irrel. bew. Dok.
M45	Anzahl richtig relevant bewerteter Dokumente	Anz. richtig rel. bew. Dok.
M46	Anzahl richtig relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente	Anz. richtig rel. bew. Dok. (erste 10 Dok.)
M47	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche	Anz. richtig rel. bew. Dok. (erste Suche)
M48	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche	Anz. richtig rel. bew. Dok. (letzte Suche)

Tab. D.3.: Durchschnittliche Bewertungen von Dokumentenmengen bei 4-stufiger Relevanzskala (Variablengrp 2). Als Dokumentenbewertung geht jeweils die letzte Bewertung eines Dokuments über alle Suchanfragen in die Betrachtung ein.

ID	Beschreibung	Kurzform
B07	Durchschnittliche Bewertung eher irrelevanter Dokumente	Durchschn. Bew. eher irrel. Dok.
B08	Durchschnittliche Bewertung eher irrelevanter Dokumente der ersten durchgeführten Suche	Durchschn. Bew. eher irrel. Dok. (erste Suche)
B09	Durchschnittliche Bewertung eher irrelevanter Dokumente der letzten durchgeführten Suche	Durchschn. Bew. eher irrel. Dok. (letzte Suche)
B10	Durchschnittliche Bewertung eher relevanter Dokumente	Durchschn. Bew. eher rel. Dok.

Fortsetzung auf nächster Seite

Tab. D.3 (Fortsetzung) Variablengrp. 2

ID	Beschreibung	Kurzform
B11	Durchschnittliche Bewertung eher relevanter Dokumente der ersten durchgeführten Suche	Durchschn. Bew. eher rel. Dok. (erste Suche)
B12	Durchschnittliche Bewertung eher relevanter Dokumente der letzten durchgeführten Suche	Durchschn. Bew. eher rel. Dok. (letzte Suche)
B13	Durchschnittliche Bewertung irrelevanter Dokumente	Durchschn. Bew. irrel. Dok.
B14	Durchschnittliche Bewertung irrelevanter Dokumente der ersten durchgeführten Suche	Durchschn. Bew. irrel. Dok. (erste Suche)
B15	Durchschnittliche Bewertung irrelevanter Dokumente der letzten durchgeführten Suche	Durchschn. Bew. irrel. Dok. (letzte Suche)
B16	Durchschnittliche Bewertung relevanter Dokumente	Durchschn. Bew. rel. Dok.
B17	Durchschnittliche Bewertung relevanter Dokumente der ersten durchgeführten Suche	Durchschn. Bew. rel. Dok. (erste Suche)
B18	Durchschnittliche Bewertung relevanter Dokumente der letzten durchgeführten Suche	Durchschn. Bew. rel. Dok. (letzte Suche)

Tab. D.4.: Durchschnittliche Betrachtungszeiten unterschiedlicher Dokumentenmengen bei 4-stufiger Relevanzskala (Variablengrp. 3). Berücksichtigt wird jeweils der erste Aufruf eines Dokuments.

ID	Beschreibung	Kurzform
Z12/Z12-log	Durchschnittliche Betrachtungszeit eher irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. eher irrel. bew. Dok.
Z13/Z13-log	Durchschnittliche Betrachtungszeit eher irrelevanter Dokumente	Durchschn. Betrachtungsz. eher irrel. Dok.
Z14/Z14-log	Durchschnittliche Betrachtungszeit eher relevant bewerteter Dokumente	Durchschn. Betrachtungsz. eher rel. bew. Dok.
Z15/Z15-log	Durchschnittliche Betrachtungszeit eher relevanter Dokumente	Durchschn. Betrachtungsz. eher rel. Dok.
Z16/Z16-log	Durchschnittliche Betrachtungszeit falsch eher irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. falsch eher irrel. bew. Dok.
Z17/Z17-log	Durchschnittliche Betrachtungszeit falsch eher relevant bewerteter Dokumente	Durchschn. Betrachtungsz. falsch eher rel. bew. Dok.
Z18/Z18-log	Durchschnittliche Betrachtungszeit falsch irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. falsch irrel. bew. Dok.
Z19/Z19-log	Durchschnittliche Betrachtungszeit falsch relevant bewerteter Dokumente	Durchschn. Betrachtungsz. falsch rel. bew. Dok.
Z20/Z20-log	Durchschnittliche Betrachtungszeit irrelevanter Dokumente	Durchschn. Betrachtungsz. irrel. Dok.
Z21/Z21-log	Durchschnittliche Betrachtungszeit irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. irrrel. bew. Dok.
Z22/Z22-log	Durchschnittliche Betrachtungszeit relevant bewerteter Dokumente	Durchschn. Betrachtungsz. rel. bew. Dok.
Z23/Z23-log	Durchschnittliche Betrachtungszeit relevanter Dokumente	Durchschn. Betrachtungsz. rel. Dok.
Z24/Z24-log	Durchschnittliche Betrachtungszeit richtig bewerteter Dokumente	Durchschn. Betrachtungsz. richtig bew. Dok.
Z25/Z25-log	Durchschnittliche Betrachtungszeit richtig eher irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. richtig eher irrel. bew. Dok.
Z26/Z26-log	Durchschnittliche Betrachtungszeit richtig eher relevant bewerteter Dokumente	Durchschn. Betrachtungsz. richtig eher rel. bew. Dok.

Fortsetzung auf nächster Seite

Tab. D.4 (Fortsetzung) Variablengrp. 3

ID	Beschreibung	Kurzform
Z27/Z27-log	Durchschnittliche Betrachtungszeit richtig irrelevant bewerteter Dokumente	Durchschn. Betrachtungsz. richtig irrel. bew. Dok.
Z28/Z28-log	Durchschnittliche Betrachtungszeit richtig relevant bewerteter Dokumente	Durchschn. Betrachtungsz. richtig rel. bew. Dok.

Tab. D.5.: Verhältnisse von Dokumentenmengen bei 4-stufiger Relevanzskala (Variablengrp. 4).

ID	Beschreibung	Kurzform	Formel
V34	Anzahl aufgerufener eher relevanter Dokumente / Anzahl eher relevanter Dokumente im Korpus	Anz. aufg. eher rel. Dok. / Anz. eher rel. Dok. im Korpus	M21 / REL05
V35	Anzahl aufgerufener eher relevanter Dokumente / Anzahl zurückgegebener eher relevanter Dokumente	Anz. aufg. eher rel. Dok. / Anz. zurückgeg. eher rel. Dok.	M21 / REL07
V36	Anzahl aufgerufener irrelevanter Dokumente / Anzahl aufgerufener Dokumente	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	M22 / M01
V37	Anzahl aufgerufener relevanter Dokumente / Anzahl aufgerufener Dokumente	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	M23 / M01
V38	Anzahl aufgerufener relevanter Dokumente / Anzahl relevanter Dokumente im Korpus	Anz. aufg. rel. Dok. / Anz. rel. Dok. im Korpus	M23 / REL06
V39	Anzahl aufgerufener relevanter Dokumente / Anzahl zurückgegebener relevanter Dokumente	Anz. aufg. rel. Dok. / Anz. zurückgeg. rel. Dok.	M23 / REL08
V40	Anzahl falsch eher irrelevant bewerteter Dokumente / Anzahl eher irrelevant bewerteter Dokumente	Anz. falsch eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	M26 / M24
V41	Anzahl falsch eher irrelevant bewerteter eher relevanter Dokumente / Anzahl eher irrelevant bewerteter Dokumente	Anz. falsch eher irrel. bew. eher rel. Dok. / Anz. eher irrel. bew. Dok.	M27 / M24
V42	Anzahl falsch eher irrelevant bewerteter irrelevanter Dokumente / Anzahl eher irrelevant bewerteter Dokumente	Anz. falsch eher irrel. bew. irrel. Dok. / Anz. eher irrel. bew. Dok.	M28 / M24
V43	Anzahl falsch eher irrelevant bewerteter relevanter Dokumente / Anzahl eher irrelevant bewerteter Dokumente	Anz. falsch eher irrel. bew. rel. Dok. / Anz. eher irrel. bew. Dok.	M29 / M24
V44	Anzahl falsch eher relevant bewerteter Dokumente / Anzahl eher relevant bewerteter Dokumente	Anz. falsch eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	M30 / M25
V45	Anzahl falsch eher relevant bewerteter eher irrelevanter Dokumente / Anzahl eher relevant bewerteter Dokumente	Anz. falsch eher rel. bew. eher irrel. Dok. / Anz. eher rel. bew. Dok.	M31 / M25
V46	Anzahl falsch eher relevant bewerteter irrelevanter Dokumente / Anzahl eher relevant bewerteter Dokumente	Anz. falsch eher rel. bew. irrel. Dok. / Anz. eher rel. bew. Dok.	M32 / M25
V47	Anzahl falsch eher relevant bewerteter relevanter Dokumente / Anzahl eher relevant bewerteter Dokumente	Anz. falsch eher rel. bew. rel. Dok. / Anz. eher rel. bew. Dok.	M33 / M25
V48	Anzahl falsch irrelevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	M34 / M01
V49	Anzahl falsch irrelevant bewerteter Dokumente / Anzahl irrelevant bewerteter Dokumente	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	M34 / M36
V50	Anzahl falsch irrelevant bewerteter Dokumente / Anzahl richtig irrelevant bewerteter Dokumente	Anz. falsch irrel. bew. Dok. / Anz. richtig irrel. bew. Dok.	M34 / M44
V51	Anzahl falsch relevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. falsch rel. bew. Dok. / Anz. aufg. Dok.	M35 / M01
V52	Anzahl falsch relevant bewerteter Dokumente / Anzahl relevant bewerteter Dokumente	Anz. falsch rel. bew. Dok. / Anz. rel. bew. Dok.	M35 / M37
V53	Anzahl falsch relevant bewerteter Dokumente / Anzahl richtig relevant bewerteter Dokumente	Anz. falsch rel. bew. Dok. / Anz. richtig rel. bew. Dok.	M35 / M45

Fortsetzung auf nächster Seite

Tab. D.5 (Fortsetzung) Variablengrp. 4

ID	Beschreibung	Kurzform	Formel
V54	Anzahl irrelevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. irrel. bew. Dok. / Anz. aufg. Dok.	M36 / M01
V55	Anzahl relevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. rel. bew. Dok. / Anz. aufg. Dok.	M37 / M01
V56	Anzahl richtig bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. richtig bew. Dok. / Anz. aufg. Dok.	M41 / M01
V57	Anzahl richtig eher irrelevant bewerteter Dokumente / Anzahl eher irrelevant bewerteter Dokumente	Anz. richtig eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	M42 / M24
V58	Anzahl richtig eher relevant bewerteter Dokumente / Anzahl eher relevant bewerteter Dokumente	Anz. richtig eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	M43 / M25
V59	Anzahl richtig eher relevant bewerteter Dokumente / Anzahl eher relevanter Dokumente im Korpus	Anz. richtig eher rel. bew. Dok. / Anz. eher rel. Dok. im Korpus	M43 / REL05
V60	Anzahl richtig eher relevant bewerteter Dokumente / Anzahl zurückgegebener eher relevanter Dokumente	Anz. richtig eher rel. bew. Dok. / Anz. zurückgeg. eher rel. Dok.	M43 / REL07
V61	Anzahl richtig irrelevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	M44 / M01
V62	Anzahl richtig irrelevant bewerteter Dokumente / Anzahl aufgerufener irrelevanter Dokumente	Anz. richtig irrel. bew. Dok. / Anz. aufg. irrel. Dok.	M44 / M22
V63	Anzahl richtig irrelevant bewerteter Dokumente / Anzahl falsch irrelevant bewerteter Dokumente	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.	M44 / M34
V64	Anzahl richtig irrelevant bewerteter Dokumente / Anzahl irrelevant bewerteter Dokumente	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	M44 / M36
V65	Anzahl richtig relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente / Anzahl aufgerufener Dokumente der ersten zehn angezeigten Dokumente	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. aufg. Dok. (erste 10 Dok.)	M46 / M02
V66	Anzahl richtig relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente / Anzahl relevant bewerteter Dokumente der ersten zehn angezeigten Dokumente	Anz. richtig rel. bew. Dok. (erste 10 Dok.) / Anz. rel. bew. Dok. (erste 10 Dok.)	M46 / M38
V67	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche / Anzahl aufgerufener Dokumente der ersten durchgeführten Suche	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. aufg. Dok. (erste Suche)	M47 / M03
V68	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche / Anzahl relevant bewerteter Dokumente der ersten durchgeführten Suche	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. bew. Dok. (erste Suche)	M47 / M39
V69	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche / Anzahl relevanter Dokumente im Korpus	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. rel. Dok. im Korpus	M47 / REL06
V70	Anzahl richtig relevant bewerteter Dokumente der ersten durchgeführten Suche / Anzahl zurückgegebener relevanter Dokumente der ersten durchgeführten Suche	Anz. richtig rel. bew. Dok. (erste Suche) / Anz. zurückgeg. rel. Dok. (erste Suche)	M47 / REL09
V71	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche / Anzahl aufgerufener Dokumente der letzten durchgeführten Suche	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. aufg. Dok. (letzte Suche)	M48 / M04
V72	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche / Anzahl relevant bewerteter Dokumente der letzten durchgeführten Suche	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. bew. Dok. (letzte Suche)	M48 / M40

Fortsetzung auf nächster Seite

Tab. D.5 (Fortsetzung) Variablengrp. 4

ID	Beschreibung	Kurzform	Formel
V73	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche / Anzahl relevanter Dokumente im Korpus	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. rel. Dok. im Korpus	M48 / REL06
V74	Anzahl richtig relevant bewerteter Dokumente der letzten durchgeführten Suche / Anzahl zurückgegebener relevanter Dokumente der letzten durchgeführten Suche	Anz. richtig rel. bew. Dok. (letzte Suche) / Anz. zurückgeg. rel. Dok. (letzte Suche)	M48 / REL10
V75	Anzahl richtig relevant bewerteter Dokumente / Anzahl aufgerufener Dokumente	Anz. richtig rel. bew. Dok. / Anz. aufg. Dok.	M45 / M01
V76	Anzahl richtig relevant bewerteter Dokumente / Anzahl aufgerufener relevanter Dokumente	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	M45 / M23
V77	Anzahl richtig relevant bewerteter Dokumente / Anzahl falsch relevant bewerteter Dokumente	Anz. richtig rel. bew. Dok. / Anz. falsch rel. bew. Dok.	M45 / M35
V78	Anzahl richtig relevant bewerteter Dokumente / Anzahl relevant bewerteter Dokumente	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.	M45 / M37
V79	Anzahl richtig relevant bewerteter Dokumente / Anzahl relevanter Dokumente im Korpus	Anz. richtig rel. bew. Dok. / Anz. rel. Dok. im Korpus	M45 / REL06
V80	Anzahl richtig relevant bewerteter Dokumente / Anzahl zurückgegebener relevanter Dokumente	Anz. richtig rel. bew. Dok. / Anz. zurückgeg. rel. Dok.	M45 / REL08

Tab. D.6.: Sonstige Leistungsmaße bei 4-stufiger Relevanzskala (Variablengrp. 5). Für S06 wird jeweils der erste Aufruf eines Dokuments berücksichtigt.

ID	Beschreibung	Kurzform
S06/S06-log	Zeit bis zum ersten richtig relevant bewerteten Dokument	Zeit zum ersten richtig rel. bew. Dok.

D.4. Benutzerleistungskovariaten zur statistischen Kontrolle des systembedingten Anpassungseffekts der Relevanzwahrnehmung

Im Folgenden wird eine Übersicht über die im dritten Experiment berücksichtigen Benutzerleistungskovariaten gegeben. Berücksichtigt werden drei Gruppen von Kovariaten: auf die Relevanzwahrnehmung bezogene Kovariaten, effektivitätsbezogene Kovariaten sowie aufwandsbezogene Kovariaten.

Tab. D.7.: Übersicht über berücksichtigte Benutzerleistungskovariaten.

ID	**Gruppe**	**Beschreibung**
B01	Wahrgenommene Qualität	Durchschn. Bew. irrel. Dok.
B02	Wahrgenommene Qualität	Durchschn. Bew. irrel. Dok. (erste Suche)
B03	Wahrgenommene Qualität	Durchschn. Bew. irrel. Dok. (letzte Suche)
B04	Wahrgenommene Qualität	Durchschn. Bew. rel. Dok.
B05	Wahrgenommene Qualität	Durchschn. Bew. rel. Dok. (erste Suche)
B06	Wahrgenommene Qualität	Durchschn. Bew. rel. Dok. (letzte Suche)
B07	Wahrgenommene Qualität	Durchschn. Bew. eher irrel. Dok.
B08	Wahrgenommene Qualität	Durchschn. Bew. eher irrel. Dok. (erste Suche)
B09	Wahrgenommene Qualität	Durchschn. Bew. eher irrel. Dok. (letzte Suche)
B10	Wahrgenommene Qualität	Durchschn. Bew. eher rel. Dok.
B11	Wahrgenommene Qualität	Durchschn. Bew. eher rel. Dok. (erste Suche)
B12	Wahrgenommene Qualität	Durchschn. Bew. eher rel. Dok. (letzte Suche)
B13	Wahrgenommene Qualität	Durchschn. Bew. irrel. Dok. (4-st.)
B14	Wahrgenommene Qualität	Durchschn. Bew. irrel. Dok. (erste Suche) (4-st.)
B15	Wahrgenommene Qualität	Durchschn. Bew. irrel. Dok. (letzte Suche) (4-st.)
B16	Wahrgenommene Qualität	Durchschn. Bew. rel. Dok. (4-st.)
B17	Wahrgenommene Qualität	Durchschn. Bew. rel. Dok. (erste Suche) (4-st.)
B18	Wahrgenommene Qualität	Durchschn. Bew. rel. Dok. (letzte Suche) (4-st.)
S05/S05-log	Effektivität	Zeit zum ersten richtig rel. bew. Dok.
S06/S06-log	Effektivität	Zeit zum ersten richtig rel. bew. Dok. (4-st.)
M10	Effektivität	Anz. rel. bew. Dok.
M16	Effektivität	Anz. richtig rel. bew. Dok.
M37	Effektivität	Anz. rel. bew. Dok. (4-st.)
M45	Effektivität	Anz. richtig rel. bew. Dok. (4-st.)
S01	Aufwand	Anz. Suchen
S02	Aufwand	Erste betr. Rankingpos.
S03	Aufwand	Letzte betr. Rankingpos.
S04	Aufwand	Suchdauer

E. Weitere Ergebnisse zu Experiment 3

Dieser Anhang enthält weitere Ergebnisse, die im Rahmen der Auswertung des dritten Experiments entstanden sind. Auf eine Darstellung innerhalb der Arbeit wird aus Gründen der Übersichtlichkeit verzichtet, es wird jedoch in vielen Fällen darauf verwiesen. Die Struktur des Kapitels ist im Wesentlichen an die Struktur von Kapitel 7 angelehnt, wobei die vertiefenderen Ergebnisse der Item- und Faktorenanalyse den weitergehenden Ergebnissen der durchgeführten Varianzanalysen vorangestellt sind. Die letzten beiden Abschnitte fassen weitere Ergebnisse in Bezug auf die Überprüfung der Gütekriterien des dritten Experiments zusammen.

E.1. Weitere Ergebnisse der Itemanalyse

Neben den in Abschnitt 7.4.4.1 berichteten Ergebnissen der Trennschärfeanalyse über alle in Experiment 3 verwendeten Zufriedenheitsitems, wird die Trennschärfe im Folgenden zusätzlich separat für die EUCS- und Zusatzitems bestimmt. Die in den Tabellen E.1 und E.2 dargestellten Trennschärfekoeffizienten unterscheiden sich jedoch kaum von den Werten der Gesamtauswertung (vgl. Abschn. 7.4.4.1).

Tab. E.1.: Trennschärfe der EUCS-Items ($n = 128$).

Item	Beschreibung	Korrigierte Item-Total-Korrelation
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	0,83
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	0,84
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	0,82
F04	Liefert die Suchmaschine genügend Information?	0,83
F05	Ist die Suchmaschine präzise?	0,85
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	0,87
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	0,77
F08	Ist die Suchmaschine benutzerfreundlich?	0,64
F09	Ist die Suchmaschine einfach zu bedienen?	0,57
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	0,75
F11	Liefert die Suchmaschine aktuelle Information?	0,55

Tab. E.2.: Trennschärfe der Zusatzsitems ($n = 128$).

Item	Beschreibung	Korrigierte Item-Total-Korrelation
F14	Es war einfach, die Aufgabe zu bearbeiten.	0,66
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	0,86
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	0,69
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	0,83
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	0,77
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	0,69
F23	Ich bin mit meiner Suchleistung zufrieden.	0,74
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	0,68

E.2. Weitere Ergebnisse der explorativen Faktorenanalyse

In diesem Abschnitt werden weitergehende Ergebnisse bezüglich der explorativen Faktorenanalysen beschrieben. Während innerhalb der Arbeit lediglich die Endresultate der einzelnen

Faktorenanalysen berichtet werden, findet sich im Folgenden ein vertiefender Einblick in die Herleitung der einzelnen Skalen. Analog zu Abschnitt 7.4.4.2 befasst sich der erste Unterabschnitt zunächst mit dem Replikationsversuch der fünf im EUCS-Instrument enthaltenen Zufriedenheitsskalen. Der dritte Unterabschnitt beschreibt die gemeinsame Analyse von EUCS- und Zusatzitems eingehender. Im letzten Unterabschnitt sind ergänzend zu den in Abschnitt 7.4.4.3 berichteten Resultaten der Reliabilitäts- und Validitätsanalyse weitere Ergebnisse in Bezug auf die in Experiment 3 betrachteten kritischen Fallgruppen angegeben.

E.2.1. Analyse der EUCS-Items

Zur Analyse der EUCS-Items werden fünf Hauptkomponentenanalysen durchgeführt, deren Ergebnisse im Folgenden dargestellt sind. Um die von Doll und Torkzadeh (1988) beschriebene Faktorenstruktur zu replizieren, wird zunächst eine Hauptkomponentenanalyse mit fünf Faktoren und Varimax-Rotation gewählt. Nach der Rotation erhält man die in E.3 dargestellte Ladungsmatrix. Dabei ergeben sich jedoch viele hohe Mehrfachladungen ($> 0{,}4$), so dass die resulierenden Komponenten nicht in eindeutiger Weise benannt werden können. Da das Auftreten von Doppel- und Mehrfachladungen als Hinweis auf eine Korreliertheit der Faktoren gewertet werden kann, wird in den folgenden Analysen eine oblique Rotationsmethode zugrunde gelegt. Da aufgrund des Ladungsmusters in Tabelle E.3 davon auszugehen ist, dass Item F07 keinen eigenen Faktor bildet, wird die Faktorenanzahl im nächsten Schritt um einen Faktor reduziert. Die Ladungsmatrix ist in Tabelle E.4 wiedergegeben. Zwar sind die meisten Mehrfachladungen jetzt verschwunden, jedoch hat sich die Interpretierbarkeit der Faktorlösung nicht wesentlich verbessert. Ferner kann die ursprüngliche Faktorstruktur auch auf diese Weise nicht repliziert werden. Deshalb wird im nächsten Schritt versucht die im zweiten Experiment gewählte Dreifaktorenlösung zu replizieren, indem zunächst erneut die Items F10 und F11 von der Analyse ausgeschlossen und anstatt vier nur noch drei Faktoren extrahiert werden. Allerdings zeigt auch das in Tabelle E.5 dargestellte Ladungsmuster nicht die erwartete Aufspaltung der ersten Komponente in die beiden Teilskalen Inhalt und Genauigkeit. Auch durch den zusätzliche Ausschluss von Item F09 lässt sich die im zweiten Experiment berichtete Dreifaktorenlösung nicht reproduzieren (vgl. Tab. E.6). Die schließlich gewählte Zweifaktorenlösung wird in Abschnitt 7.4.4.2 erläutert.

Tab. E.3.: PCA 1: EUCS-Items mit 5 Faktoren und Varimax-Rotation (alle 11 EUCS-Items). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,86**	0,19	0,26	0,19	0,06
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,85**	0,18	0,24	0,07	**0,30**
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,82**	**0,31**	0,12	0,20	0,18
F04	Liefert die Suchmaschine genügend Information?	**0,72**	**0,41**	0,10	0,22	0,27
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,67**	**0,52**	0,11	0,21	**0,32**
F05	Ist die Suchmaschine präzise?	**0,57**	**0,53**	0,15	**0,32**	**0,33**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	**0,48**	**0,74**	0,22	0,11	0,12
F09	Ist die Suchmaschine einfach zu bedienen?	0,20	0,06	**0,88**	0,23	0,23
F08	Ist die Suchmaschine benutzerfreundlich?	0,21	**0,59**	**0,71**	0,06	0,05
F11	Liefert die Suchmaschine aktuelle Information?	0,24	0,14	0,21	**0,92**	0,12
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	**0,44**	0,19	**0,34**	0,17	**0,76**

Tab. E.4.: PCA 2: EUCS-Items mit 4 Faktoren und Oblimin-Rotation (alle 11 EUCS-Items). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**1,00**	0,00	−0,11	−0,16
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,92**	−0,08	0,06	0,05
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,87**	0,00	0,04	−0,04
F04	Liefert die Suchmaschine genügend Information?	**0,86**	−0,03	0,09	0,10
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,84**	0,05	0,067	0,16
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	**0,68**	0,23	0,04	**-0,38**
F05	Ist die Suchmaschine präzise?	**0,68**	0,11	0,21	0,17
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	**0,54**	**0,35**	0,00	**0,43**
F08	Ist die Suchmaschine benutzerfreundlich?	0,05	**0,91**	−0,03	0,22
F09	Ist die Suchmaschine einfach zu bedienen?	−0,01	**0,81**	0,16	**-0,34**
F11	Liefert die Suchmaschine aktuelle Information?	0,02	0,00	**0,98**	0,00

Tab. E.5.: PCA 3: EUCS-Items mit 3 Faktoren und Oblimin-Rotation (EUCS-Items ohne F10 u. F11). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,95**	−0,04	−0,04
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,94**	0,09	−0,19
F04	Liefert die Suchmaschine genügend Information?	**0,90**	−0,05	0,07
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,89**	0,08	−0,12
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,87**	−0,03	0,18
F05	Ist die Suchmaschine präzise?	**0,78**	0,07	0,19
F09	Ist die Suchmaschine einfach zu bedienen?	0,03	**0,98**	−0,10
F08	Ist die Suchmaschine benutzerfreundlich?	0,05	**0,65**	**0,51**
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	**0,56**	0,03	**0,56**

Tab. E.6.: PCA 4: EUCS-Items mit 3 Faktoren und Oblimin-Rotation (EUCS-Items ohne F10, F11 u. F09). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3
F05	Ist die Suchmaschine präzise?	**0,89**	0,06	−0,01
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	**0,88**	−0,15	0,24
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	**0,74**	0,25	0,00
F04	Liefert die Suchmaschine genügend Information?	**0,64**	**0,36**	−0,07
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	−0,04	**0,91**	0,15
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	0,13	**0,83**	0,03
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,44**	**0,56**	−0,07
F08	Ist die Suchmaschine benutzerfreundlich?	0,05	0,09	**0,92**

E.2.2. Analyse aller Items

Um zu zeigen, dass die im zweiten Experiment gefundene Faktorenstruktur ein valides Bild der untersuchten Dimensionen der Benutzerzufriedenheit wiedergibt, wird zunächst eine Hauptkomponentenanalyse mit den in der Lösung von 2009 enthaltenen Items und gleicher Faktorenanzahl durchgeführt. Das Ladungsmuster dieser Lösung ist Tabelle E.7 zu entnehmen. Es lässt erkennen, dass sich die Daten des dritten Experiments inhaltlich zu den gleichen Faktoren zusammenfassen lassen wie im zweiten Experiment. Jedoch laden insbesondere die Items F19 und F23 nicht eindeutig auf einem Faktor. Da für diese erste Lösung zudem das Multikollinearitätsproblem unberücksichtigt gelassen wurde, wird im nächsten Schritt geprüft, welche Items zusätzlich entfernt

werden müssen, damit Multikollinearität ausgeschlossen werden kann. Es zeigt sich, dass die Determinante der Korrelationsmatrix durch einen Ausschluss der Items F02 und F16 über dem kritischen Schwellenwert liegt.

Tab. E.7.: PCA 1: Alle Items mit 4 Faktoren und Oblimin-Rotation (alle 15 Items). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,90**	−0,08	0,09	0,02
F04	Liefert die Suchmaschine genügend Information?	**0,89**	0,00	0,01	0,00
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,87**	−0,03	0,05	0,06
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	**0,78**	−0,06	0,20	0,09
F05	Ist die Suchmaschine präzise?	**0,68**	**0,32**	−0,06	0,05
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	**0,51**	**0,40**	0,14	−0,22
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,51**	0,22	0,18	0,21
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	−0,09	**0,84**	0,17	0,07
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	−0,02	**0,75**	0,09	0,21
F08	Ist die Suchmaschine benutzerfreundlich?	**0,31**	**0,61**	0,02	**-0,38**
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,49**	**0,53**	−0,09	0,06
F14	Es war einfach, die Aufgabe zu bearbeiten.	0,01	0,01	**0,95**	−0,04
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	**0,35**	0,17	**0,51**	0,12
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	**0,36**	0,18	0,00	**0,63**
F23	Ich bin mit meiner Suchleistung zufrieden.	0,20	0,14	**0,38**	**0,44**

Deshalb wird im nächsten Schritt eine Hauptkomponentenanalyse mit den verbleibenden 13 Items gerechnet. Das Ergebnis dieser Analyse ist in Tabelle E.8 dokumentiert. Es bleibt festzustellen, dass die Faktorenstruktur des zweiten Experiments nach wie vor zu erkennen ist. Außerdem fällt auf, dass insbesondere die Items F08 und F23 nicht besonders eindeutig nur auf einem der vier Faktoren laden. Dies führt dazu, dass Item F08, als Item mit der höchsten Anzahl an Nebenladungen, für die in Abschnitt 7.4.4.2 berichtete Lösung aus der Berechnung entfernt wird.

Tab. E.8.: PCA 2: Alle Items mit 4 Faktoren und Oblimin-Rotation (ohne F02 u. F16). Die den jeweiligen Faktor charakterisierenden Faktorladungen sind fett hervorgehoben.

Item	Beschreibung	Faktor 1	Faktor 2	Faktor 3	Faktor 4
F04	Liefert die Suchmaschine genügend Information?	**0,92**	−0,07	0,04	0,02
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	**0,89**	−0,13	0,13	0,05
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**0,87**	−0,07	0,07	0,08
F05	Ist die Suchmaschine präzise?	**0,73**	0,23	−0,02	0,03
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**0,62**	**0,42**	−0,12	0,03
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	**0,61**	0,29	0,13	−0,24
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	**0,56**	0,19	0,17	0,20
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	−0,01	**0,82**	0,14	0,02
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	−0,01	**0,79**	0,10	0,17
F08	Ist die Suchmaschine benutzerfreundlich?	**0,39**	**0,51**	0,01	**–0,40**
F14	Es war einfach, die Aufgabe zu bearbeiten.	0,02	0,02	**0,96**	−0,02
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	**0,36**	0,22	0,01	**0,62**
F23	Ich bin mit meiner Suchleistung zufrieden.	0,22	0,20	**0,33**	**0,45**

E.2.3. Reliabilitäts- und Validitätsanalyse

Im Folgenden sind die Auswertung der Gütekriterien der im dritten Experiment ermittelten Skalen nach Ausschluss der kritischen Fallgruppen SP_{SB} (Suchbegriff mehrdeutig), SP_{MV} (Manipulation versagt) und SP_{TD}(Test durchschaut) sowie die Analysen der Originalskalen des EUCS-Instruments dargestellt. Die Tabellen E.9 bis E.11 zeigen zunächst, die Reliabilitäts- und Validitätskoeffizienten der im dritten Experiment ermittelten Skalen für die Gesamtstichprobe, nachdem jeweils eine der oben genannten Fallgruppen ausgeschlossen wurde. Die Überprüfung

der Standardabweichung zeigt auch hier, dass die ermittelten Koeffizienten im Vergleich stabil bleiben und somit die Möglichkeit einer Beeinflussung der Ergebnisse durch die Hinzunahme der hier ausgeschlossenen Fälle eher gering ist.

Tab. E.9.: Skalenreliabilität und Kriteriumsvalidität unter Ausschluss kritischer Fallgruppen für $A1$. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	**Cronbachs Alpha**			**Kriteriumsvalidität**		
	$\mathbf{SP_A \setminus SP_{SB}}$	$\mathbf{SP_A \setminus SP_{MV}}$	$\mathbf{SP_A \setminus SP_{TD}}$	$\mathbf{SP_A \setminus SP_{SB}}$	$\mathbf{SP_A \setminus SP_{MV}}$	$\mathbf{SP_A \setminus SP_{TD}}$
	$n = 95$	$n = 120$	$n = 123$	$n = 95$	$n = 120$	$n = 123$
SK12	0,9	0,89	0,89	0,79	0,81	0,79
SK13	0,57	0,57	0,55	0,57	0,54	0,55
SK14	0,84	0,79	0,8	0,77	0,76	0,75
SK15/19	0,81	0,76	0,73	0,64	0,6	0,6
SK16	0,76	0,78	0,79	0,65	0,66	0,66
SK17	0,89	0,89	0,89	0,81	0,83	0,81
SK18	0,69	0,68	0,67	0,67	0,61	0,6

Tab. E.10.: Skalenreliabilität und Kriteriumsvalidität unter Ausschluss kritischer Fallgruppen für $A2$. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	**Cronbachs Alpha**			**Kriteriumsvalidität**		
	$\mathbf{SP_A \setminus SP_{SB}}$	$\mathbf{SP_A \setminus SP_{MV}}$	$\mathbf{SP_A \setminus SP_{TD}}$	$\mathbf{SP_A \setminus SP_{SB}}$	$\mathbf{SP_A \setminus SP_{MV}}$	$\mathbf{SP_A \setminus SP_{TD}}$
	$n = 95$	$n = 120$	$n = 123$	$n = 95$	$n = 120$	$n = 123$
SK12	0,95	0,95	0,95	0,89	0,89	0,90
SK13	0,78	0,76	0,74	0,63	0,54	0,56
SK14	0,84	0,87	0,86	0,81	0,8	0,8
SK15/19	0,82	0,81	0,81	0,76	0,76	0,76
SK16	0,89	0,9	0,89	0,78	0,78	0,78
SK17	0,94	0,95	0,95	0,9	0,89	0,89
SK18	0,74	0,74	0,75	0,72	0,72	0,73

Tab. E.11.: Skalenreliabilität und Kriteriumsvalidität unter Ausschluss kritischer Fallgruppen für $A3$. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	**Cronbachs Alpha**			**Kriteriumsvalidität**		
	$\mathbf{SP_A \setminus SP_{SB}}$	$\mathbf{SP_A \setminus SP_{MV}}$	$\mathbf{SP_A \setminus SP_{TD}}$	$\mathbf{SP_A \setminus SP_{SB}}$	$\mathbf{SP_A \setminus SP_{MV}}$	$\mathbf{SP_A \setminus SP_{TD}}$
	$n = 95$	$n = 120$	$n = 123$	$n = 95$	$n = 120$	$n = 123$
SK12	0,96	0,96	0,96	0,90	0,91	0,92
SK13	0,72	0,72	0,73	0,66	0,62	0,63
SK14	0,87	0,88	0,88	0,87	0,87	0,87
SK15/19	0,81	0,81	0,81	0,84	0,81	0,83
SK16	0,88	0,88	0,88	0,83	0,83	0,85
SK17	0,95	0,95	0,95	0,89	0,9	0,91
SK18	0,8	0,76	0,77	0,82	0,81	0,83

Die Reliabilitäts- und Validitätskoeffizienten der Originalskalen des EUCS-Instruments sind in den Tabellen E.12 und E.13 zusammengefasst. Die Überprüfung der Gütekriterien weist wie im zweiten Experiment darauf hin, dass bei den Skalen *Ease of use* und *Timeliness* sichtbare Optimierungsbedarfe bestehen. Auch im dritten Experiment wird durch die nach Aufgaben getrennte Analyse weiterhin deutlich, dass die zu diesen beiden Skalen beitragenden Frageitems scheinbar einer gewissen Einarbeitungsphase bedürfen, um zuverlässig beantwortet werden zu können. Dies erklärt sich aus den höheren internen Konsistenzen der zweiten und dritten Aufgabe. Für die Skalen *Content* und *Accuracy* sind die internen Konsistenzen erneut in allen fünf Fallgruppen zufriedenstellend ($\alpha > 0{,}7$). Darüber hinaus wird die Kriteriumsvalidität für alle vier Skalen

bestätigt. In einem weiteren Auswertungsschritt findet der Vergleich zwischen den verschiedenen Fallgruppen statt. Hier zeigt sich erneut, dass die Abweichungen zwischen den betrachteten Fallgruppen minimal sind.

Tab. E.12.: Skalenreliabilität der Originalskalen des EUCS-Instruments nach Datenqualität und unter Ausschluss kritischer Fallgruppen. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	**Cronbachs Alpha**				
	SP_A	**SP_B**	**$SP_A \setminus SP_{SB}$**	**$SP_A \setminus SP_{MV}$**	**$SP_A \setminus SP_{TD}$**
	$n = 128$	$n = 86$	$n = 95$	$n = 120$	$n = 123$
Aufgabe 1					
Content	0,86	0,87	0,87	0,86	0,86
Accuracy	0,85	0,87	0,87	0,85	0,85
Ease of use	0,57	0,54	0,57	0,57	0,55
Timeliness	0,43	0,51	0,48	0,46	0,42
Aufgabe 2					
Content	0,93	0,92	0,92	0,93	0,93
Accuracy	0,91	0,91	0,91	0,91	0,91
Ease of use	0,75	0,76	0,78	0,76	0,74
Timeliness	0,67	0,67	0,68	0,66	0,67
Aufgabe 3					
Content	0,95	0,95	0,94	0,95	0,95
Accuracy	0,91	0,91	0,91	0,9	0,91
Ease of use	0,73	0,72	0,72	0,72	0,73
Timeliness	0,7	0,73	0,73	0,68	0,7

Tab. E.13.: Kriteriumsvalidität der Originalskalen des EUCS-Instruments nach Datenqualität und unter Ausschluss kritischer Fallgruppen. Aus SP_A ausgeschlossen werden die Fallgruppen SP_{SB}, SP_{MV} und SP_{TD}.

Skala	**Kriteriumsvalidität**				
	SP_A $n = 128$	**SP_B** $n = 86$	**$SP_A \setminus SP_{SB}$** $n = 95$	**$SP_A \setminus SP_{MV}$** $n = 120$	**$SP_A \setminus SP_{TD}$** $n = 123$
Aufgabe 1					
Content	0,73	0,7	0,7	0,74	0,73
Accuracy	0,75	0,79	0,79	0,78	0,75
Ease of use	0,55	0,58	0,57	0,54	0,55
Timeliness	0,61	0,57	0,59	0,61	0,6
Aufgabe 2					
Content	0,86	0,87	0,87	0,86	0,86
Accuracy	0,86	0,86	0,85	0,86	0,87
Ease of use	0,56	0,63	0,63	0,54	0,56
Timeliness	0,8	0,79	0,79	0,79	0,8
Aufgabe 3					
Content	0,89	0,89	0,88	0,89	0,9
Accuracy	0,89	0,89	0,9	0,88	0,88
Ease of use	0,63	0,66	0,66	0,62	0,63
Timeliness	0,77	0,76	0,75	0,77	0,78

E.3. Weitere Ergebnisse der Varianzanalysen

In diesem Anhang sind weitere Ergebnisse zu den im Rahmen des dritten Experiments durchgeführten Varianzanalysen zusammengestellt auf deren Darstellung innerhalb der Arbeit aus Gründen der Übersichtlichkeit verzichtet wird. Zur besseren Übersicht ist der vorliegende Abschnitt in fünf Teile gegliedert. Abschnitt E.3.1 enthält zunächst eine Übersicht über diejenigen Variablen die in keiner der fünf Stichproben einen signifikanten Effekt der untersuchten Einflussgrößen zeigen. Abschnitt E.3.2 umfasst die Mittelwerte der Interaktionen für die in den Abschnitten 7.4.3.1 und 7.4.5.1 berichteten signifikanten Ergebnisse der Varianzanalysen. Im Anschluss daran geben die in Abschnitt E.3.3 dargestellten Tabellen Aufschluss über das im Einzelnen verwendete Analyseverfahren, die Stabilität der berichteten Effekte (eindeutig vs. tendenziell) sowie das Signifikanzniveau der jeweiligen unabhängigen Variablen.

E.3.1. Variablen ohne signifikante Unterschiede

Tabelle E.14 stellt eine Übersicht derjenigen Variablen bereit, die in keiner der fünf Stichproben einen signifikanten Effekt der untersuchten Einflussgrößen zeigen. Das Fehlen eines Effekts wird aufgeschlüsselt nach den beiden Datenqualitätsstufen SP_A und SP_B. Darüber hinaus gibt die Tabelle die Stichprobengröße sowie die Art der Varianzanalyse (klassisch vs. robust) an. Minuszeichen ($-$) bedeuten, dass für diese Teilstichprobe keine ausreichende Fallanzahl für eine Analyse zur Verfügung steht, Pluszeichen hingegen ($+$), dass signifikante Effekte der unabhängigen Variable vorliegen. Des Weiteren kennzeichnen Sterne ($*$) Variablen, für die keine eindeutige Aussage getroffen werden kann, da zwischen einer und drei Zufallsstichproben signifikante Abhängigkeiten aufweisen.

Tab. E.14.: Übersicht über Variablen ohne signifikante Unterschiede.

		SP_A		**SP_B**	
ID	**Beschreibung**	**n**	**V**	**n**	**V**
M01	Anz. aufg. Dok.	116	R	80	R

[a] Nur einmal robuste Analyse.

[b] Nur einmal klassische Analyse.

Fortsetzung auf nächster Seite

Tab. E.14 (Fortsetzung)

		SP_A		SP_B	
ID	**Beschreibung**	**n**	**V**	**n**	**V**
M03	Anz. aufg. Dok. (erste Suche)	116	R	80	R
M04	Anz. aufg. Dok. (letzte Suche)	116	R	80	R
M06	Anz. aufg. rel. Dok.	116	R	80	R
M09	Anz. irrel. bew. Dok.	116	R	80	R
M10	Anz. rel. bew. Dok.	116	R	29	R
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	20	R	80	R
M12	Anz. rel. bew. Dok. (erste Suche)	116	R	29	R
M13	Anz. rel. bew. Dok. (letzte Suche)	116	R	80	R
M16	Anz. richtig rel. bew. Dok.	116	R	29	R
M18	Anz. richtig rel. bew. Dok. (erste Suche)	116	R	80	R
M21	Anz. aufg. eher rel. Dok.	20	R	80	R
M23	Anz. aufg. rel. Dok. (4-st.)	20	K	80	K
M24	Anz. eher irrel. bew. Dok.	116	R	80	R
M25	Anz. eher rel. bew. Dok.	116	R	80	R
M28	Anz. falsch eher irrel. bew. irrel. Dok.	116	R	80	R
M29	Anz. falsch eher irrel. bew. rel. Dok.	20	R	80	R
M30	Anz. falsch eher rel. bew. Dok.	116	R	80	R
M32	Anz. falsch eher rel. bew. irrel. Dok.	116	R	80	R
M33	Anz. falsch eher rel. bew. rel. Dok.	116	R	80	R
M34	Anz. falsch irrel. bew. Dok. (4-st.)	116	R	80	R
M35	Anz. falsch rel. bew. Dok. (4-st.)	116	R	80	R
M36	Anz. irrel. bew. Dok. (4-st.)	116	R	80	R
M37	Anz. rel. bew. Dok. (4-st.)	116	R	29	R
M38	Anz. rel. bew. Dok. (erste 10 Dok.) (4-st.)	116	R	80	R
M39	Anz. rel. bew. Dok. (erste Suche) (4-st.)	116	R	80	R
M40	Anz. rel. bew. Dok. (letzte Suche) (4-st.)	116	R	29	R
M41	Anz. richtig bew. Dok. (4-st.)	116	R	80	R
M43	Anz. richtig eher rel. bew. Dok.	116	R	80	R
M45	Anz. richtig rel. bew. Dok. (4-st.)	116	R	80	K/R
M46	Anz. richtig rel. bew. Dok. (erste 10 Dok.) (4-st.)	116	R	80	R
M47	Anz. richtig rel. bew. Dok. (erste Suche) (4-st.)	116	R	80	R
M48	Anz. richtig rel. bew. Dok. (letzte Suche) (4-st.)	116	R	29	R
B05	Durchschn. Bew. rel. Dok. (erste Suche)	–	–	48	K/R[b]
B18	Durchschn. Bew. rel. Dok. (letzte Suche) (4-st.)	84	K♭R	+	+
Z01	Durchschn. Betrachtungsz. aller Dok.	116	R	+	+
Z01-log	Durchschn. Betrachtungsz. aller Dok.	116	R	+	+
Z07	Durchschn. Betrachtungsz. rel. bew. Dok.	108	R	27	R
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	108	R	72	K
Z08	Durchschn. Betrachtungsz. rel. Dok.	116	R	29	R
Z08-log	Durchschn. Betrachtungsz. rel. Dok.	116	R	29	R
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	112	R	+	+
Z09-log	Durchschn. Betrachtungsz. richtig bew. Dok.	112	K	28	K
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	108	R	+	+
Z11-log	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	18	K	72	K
Z14	Durchschn. Betrachtungsz. eher rel. bew. Dok.	–	–	40	K/R
Z22	Durchschn. Betrachtungsz. rel. bew. Dok. (4-st.)	80	R	20	R
Z22-log	Durchschn. Betrachtungsz. rel. bew. Dok. (4-st.)	80	K	52	K/R
Z23	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	116	R	29	R
Z23-log	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	116	R	29	R
Z24	Durchschn. Betrachtungsz. richtig bew. Dok. (4-st.)	88	K♭R	22	K♭R

[a] Nur einmal robuste Analyse.

[b] Nur einmal klassische Analyse.

Fortsetzung auf nächster Seite

Tab. E.14 (Fortsetzung)

		SP_A		SP_B	
ID	**Beschreibung**	**n**	**V**	**n**	**V**
Z24-log	Durchschn. Betrachtungsz. richtig bew. Dok. (4-st.)	88	K/R[b]	22	K/R[b]
V03	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	116	R	29	R
V04	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	116	R	29	R
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	116	K	+	+
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	116	K/R[b]	+	+
V24	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. aufg. Dok. (letzte Suche)}}$	80	K	52	K
V26	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. rel. Dok. im Korpus}}$	80	R	20	R
V27	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. zurückgeg. rel. Dok. (letzte Suche)}}$	80	R	20	R
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	108	R	72	K
V32/BR	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	108	R	72	R
V33	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	108	R	+	+
V34	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$	116	R	80	R
V35	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	116	R	80	R
V38	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$ (4-st.)	116	K	76	R
V39	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$ (4-st.)	116	R	76	R
V41	$\frac{\text{Anz. falsch eher irrel. bew. eher rel. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	52	R	+	+
V42	$\frac{\text{Anz. falsch eher irrel. bew. irrel. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	52	R	36	R
V43	$\frac{\text{Anz. falsch eher irrel. bew. rel. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	9	K/R[b]	36	K/R[b]
V46	$\frac{\text{Anz. falsch eher rel. bew. irrel. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	10	R	40	R
V47	$\frac{\text{Anz. falsch eher rel. bew. rel. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	68	K	40	K/R[b]
V48	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	116	R	29	R
V49	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$ (4-st.)	36	K/R[b]	9	K/R[b]
V51	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	116	R	80	R
V52	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$ (4-st.)	80	R	52	K/R
V54	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	116	R	+	+
V55	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	116	K	80	K
V56	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	116	R	80	R
V58	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	+	+	40	K
V59	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$	116	R	80	R
V60	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	116	R	80	R
V64	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$ (4-st.)	7	K/R[b]	28	K/R[b]
V78	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$ (4-st.)	56	K	32	K/R
V79	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$ (4-st.)	56	K/R	32	K/R

[a] Nur einmal robuste Analyse.

[b] Nur einmal klassische Analyse.

Fortsetzung auf nächster Seite

Tab. E.14 (Fortsetzung)

		SP_A		SP_B	
ID	**Beschreibung**	**n**	**V**	**n**	**V**
V80	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$ (4-st.)	56	K♭R	32	K♭R
S01	Anz. Suchen	116	R	80	R
S02	Erste betr. Rankingpos.	116	R	80	R
S03	Letzte betr. Rankingpos.	116	R	80	R
S04	Suchdauer	116	R	80	R
S05-log	Zeit zum ersten richtig rel. bew. Dok.	17	K	68	K
S05	Zeit zum ersten richtig rel. bew. Dok.	104	R	26	R

[a] Nur einmal robuste Analyse.
[b] Nur einmal klassische Analyse.

E.3.2. Mittelwerte der Interaktionen

Wie schon im Kontext des zweiten Experiments, enthalten die Tabellen im Hauptteil der Arbeit lediglich die nach Erwartungshaltung und Systemgüte getrennt berechneten Gruppenmittelwerte. Der Vollständigkeit halber sind diese Tabellen im folgenden Abschnitt noch einmal, um die Gruppenmittelwerte der vier unterschiedlichen Untersuchungsgruppen ergänzt, aufgeführt.

Tab. E.15.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei binärer Relevanzskala in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

ID	Beschreibung	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
M02[a]	Anz. aufg. Dok. (erste 10 Dok.)	3,08	2,80	2,90	2,99	**3,47**	**2,84**	**2,35**	**3,30**
M05[a]	Anz. aufg. irrel. Dok.	**1,5**[b]	**2,69**	1,95	2,24	1,49	1,51	2,41	2,97
M07[a]	Anz. falsch irrel. bew. Dok.	**2,38**	**1,65**[b]	1,98	2,05	2,51	2,26	1,45	1,84
M08[a]	Anz. falsch rel. bew. Dok.	**0,39**[b]	**0,72**	0,57	0,54	0,46	0,32	0,68	0,76
M14[a]	Anz. richtig bew. Dok.	**3,84**	**5,86**[b]	4,66	5,04	3,77	3,91	5,54	6,17
M15[a]	Anz. richtig irrel. bew. Dok.	**1,14**	**2,12**[b]	1,49	1,77	1,03	1,24	1,94	2,29
B04	Durchschn. Bew. rel. Dok.	**5,29**	**5,73**[b]	5,55	5,47	5,3	5,27	5,79	5,66
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	**5,36**	**5,82**[b]	5,68	5,5	5,35	5,36	6,00	5,64
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	**42,14**	**37,61**[b]	41,4	38,35	44,77	39,5	38,02	37,2
Z02-log[a]	Durchschn. Betrachtungsz. falsch bew. Dok.	**3,48**	**3,33**[b]	3,46	3,35	3,54	3,41	3,37	3,29
Z05[a]	Durchschn. Betrachtungsz. irrel. bew. Dok.	**37,42**	**32,99**[b]	36,75	33,66	38,4	36,44	35,11	30,87
Z05-log[a]	Durchschn. Betrachtungsz. irrel. bew. Dok.	**3,33**	**3,18**[b]	3,28	3,23	3,33	3,33	3,23	3,12
V01[a]	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,15**[b]	**0,28**	0,21	0,22	0,14	0,16	0,27	0,29
V02[a]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,85**[b]	**0,72**	0,8	0,77	0,86	0,84	0,73	0,71
V05[a]	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,3**	**0,19**[b]	0,23	0,25	0,3	0,29	0,16	0,21
V06[a]	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	**0,69**	**0,44**[b]	0,6	0,54	0,73	0,65	0,47	0,42
V08[a]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,04**[b]	**0,08**	0,06	0,05	0,04	0,03	0,08	0,07
V09[a]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	**0,07**[b]	**0,12**	0,1	0,09	0,07	0,06	0,12	0,11
V10[a]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. richtig rel. bew. Dok.}}$	**0,09**[b]	**0,17**	0,13	0,13	0,09	0,09	0,16	0,17
V13[a]	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,4**	**0,59**[b]	0,47	0,52	0,35	0,44	0,58	0,6
V14[a]	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,11** **0,11**	**0,21**[b] **0,21**[b]	**0,14** **0,14**	**0,18**[b] **0,18**[b]	0,09 0,09	0,12 0,13	0,19 0,19	0,23 0,22
V17[a]	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	**0,32**	**0,57**[b]	0,41	0,49	0,27	0,37	0,54	0,61
V31/BP[a]	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	**0,94**[b]	**0,9**	0,92	0,92	0,93	0,95	0,9	0,89

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.

[b] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Tab. E.16.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei binärer Relevanzskala in SP_B. Neben den in SP_B neu hinzukommenden Effekten beinhaltet diese Tabelle auch die Gruppenmittelwerte der Effekte, die aus SP_Abestätigt werden. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
M02	Anz. aufg. Dok. (erste 10 Dok.)	3,14	3,01	3,28	2,86	**3,83**	**2,47**	**2,69**	**3,36**
M05	Anz. aufg. irrel. Dok.	**1,48**[b]	**3,03**	2,37	2,14	1,58	1,38	3,15	2,9
M07	Anz. falsch irrel. bew. Dok.	**2,55**	**1,65**[b]	2,04	2,16	2,68	2,42	1,4	1,9
M08	Anz. falsch rel. bew. Dok.	**0,38**[b]	**0,81**	0,68	0,51	0,45	0,3	0,9	0,72
M14	Anz. richtig bew. Dok.	**4**	**6,33**[b]	5,38	4,96	4,07	3,93	6,68	5,98
M15	Anz. richtig irrel. bew. Dok.	**1,11**	**2,38**[b]	1,7	1,78	1,13	1,08	2,28	2,47
M17[a]	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	1,75	1,43	1,68	1,43	**2,11**	**1,31**	**1,33**	**1,53**
M19[a]	Anz. richtig rel. bew. Dok. (letzte Suche)	3,34	3,73	**4,06**[b]	**3,01**	3,88	2,8	4,23	3,23
B04[a]	Durchschn. Bew. rel. Dok.	5,33	5,56	**5,68**[b]	**5,21**	5,47	5,19	5,89	5,22
B06[a]	Durchschn. Bew. rel. Dok. (letzte Suche)	5,33	5,69	**5,81**[b]	**5,21**	5,53	5,13	6,08	5,29
Z01[a]	Durchschn. Betrachtungsz. aller Dok.	**54,64**	**40,27**[b]	44,25	50,65	49,44	59,83	39,06	41,47
Z01-log[a]	Durchschn. Betrachtungsz. aller Dok.	**3,64**	**3,44**[b]	3,47	3,61	3,56	3,72	3,38	3,49
Z02-log	Durchschn. Betrachtungsz. falsch bew. Dok.	**3,45**	**3,25**[b]	3,39	3,3	3,51	3,38	3,28	3,22
Z05[a]	Durchschn. Betrachtungsz. irrel. bew. Dok.	**37,85** **37,06**	**26,36**[b] **30,26**[b]	31,35 33,84	32,41 33,48	**42,11** **40,45**	**33,26** **33,66**	**23,12** **27,22**	**30,70** **33,3**
Z05-log	Durchschn. Betrachtungsz. irrel. bew. Dok.	**3,34**	**3,09**[b]	3,23	3,21	3,41	3,27	3,04	3,14
Z09[a]	Durchschn. Betrachtungsz. richtig bew. Dok.	**56,91**	**44,74**[b]	47,82	53,84	50,58	63,24	45,05	44,43
Z11[a]	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	**64,73**	**49,42**[b]	53,39	60,76	56,98	72,48	49,8	49,03
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,16**[b]	**0,29**	0,22	0,22	0,15	0,16	0,29	0,29
V02	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,85**[b]	**0,71**	0,77	0,77	0,85	0,84	0,7	0,71
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,3**	**0,18**[b]	0,21	0,27	0,29	0,3	0,13	0,23
V06	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	**0,7**	**0,42**[b]	0,57	0,54	0,74	0,65	0,41	0,42
V08[a]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,05**[b]	**0,08**	**0,08**	**0,05**[b]	0,05	0,04	0,1	0,06
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,05**[b]	**0,07**	**0,07**	**0,05**[b]	0,05	0,04	0,09	0,05
V09	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	**0,07**[b]	**0,12**	0,1	0,08	0,06	0,07	0,14	0,09
V10	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. richtig rel. bew. Dok.}}$	**0,09**[b]	**0,16**	0,12	0,12	0,07	0,1	0,17	0,14
V11[a]	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	0,41	0,41	**0,37**[b]	**0,45**	0,39	0,43	0,34	0,47
V12[a]	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	0,59	0,58	**0,63**[b]	**0,54**	0,61	0,57	0,65	0,51
V13	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,4**	**0,6**[b]	0,48	0,52	0,37	0,43	0,58	0,61
V14[a]	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,11**	**0,23**[b]	0,16	0,19	0,1	0,12	0,21	0,25

[a] Effekt kommt in SP_B neu hinzu.

Fortsetzung auf nächster Seite

[b] Dieser Mittelwert entspricht der besseren Benutzerleistung.

[c] Stichprobengröße < 40.

Tab. E.16 (Fortsetzung)

ID	Beschreibung	System		Erwartung		Interaktion			
		S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
V16[a]	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.	**0,51**	**1,08**[b]	0,87	0,72	0,43	0,59	1,3	0,85
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	**0,31**	**0,55**[b]	0,42	0,44	0,26	0,35	0,57	0,53
V29[a]	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	**0,66**	**0,77**[b]	0,74	0,69	0,66	0,66	0,81	0,72
V31/BP	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.	**0,94**[b]	**0,89**	0,91	0,92	0,94	0,93	0,87	0,91
V33[a]	Anz. richtig rel. bew. Dok. / Anz. zurückgeg. rel. Dok.	**0,08**	**0,11**[b]	0,1	0,09	0,09	0,07	0,11	0,1

[a] Effekt kommt in SP_B neu hinzu.
[b] Dieser Mittelwert entspricht der besseren Benutzerleistung.
[c] Stichprobengröße < 40.

Tab. E.17.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei 4-stufiger Relevanzskala in SP_B. Neben den in SP_B neu hinzukommenden Effekten beinhaltet diese Tabelle auch die Gruppenmittelwerte der Effekte, die aus SP_A bestätigt werden. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

ID	Beschreibung	System		Erwartung		Interaktion			
		S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
M20	Anz. aufg. eher irrel. Dok.	**0,93**[b]	**1,83**	1,48	1,28	0,92	0,93	2,03	1,63
M22	Anz. aufg. irrel. Dok.	**0,56**[b]	**1,33**	0,96	0,93	0,67	0,45	1,25	1,4
M27[a]	Anz. falsch eher irrel. bew. eher rel. Dok.	0,68	0,47	0,61	0,54	**0,86**	**0,42**	**0,28**	**0,56**
M31	Anz. falsch eher rel. bew. eher irrel. Dok.	**0,17**[b]	**0,41**	0,35	0,23	0,22	0,12	0,48	0,33
M42	Anz. richtig eher irrel. bew. Dok.	**0,2**	**0,49**[b]	0,38	0,31	0,18	0,22	0,57	0,4
M44	Anz. richtig irrel. bew. Dok.	**0,39**	**0,91**[b]	0,63	0,67	0,47	0,3	0,78	1,03
B16[a]	Durchschn. Bew. rel. Dok.	5,74	6,03	**6,18**[b]	**5,59**	5,94	5,53	6,41	5,65
B18[a]	Durchschn. Bew. rel. Dok. (letzte Suche)	5,69	6,01	**6,24**[b]	**5,45**	6	5,37	6,48	5,53
Z15[a]	Durchschn. Betrachtungsz. eher rel. Dok.	**50,99**	**35,64**[b]	43,02	43,61	50,23	51,74	35,81	35,47
V36	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	**0,06**[b]	**0,12**	0,08	0,09	0,06	0,05	0,1	0,13
V40[c]	Anz. falsch eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	**0,87**	**0,7** [b]	0,77	0,82	**0,93**	**0,79**	**0,59**	**0,84**
V41[a,c]	Anz. falsch eher irrel. bew. eher rel. Dok. / Anz. eher irrel. bew. Dok.	**0,44**	**0,24**[b]	0,42	0,26	0,58	0,31	0,26	0,21
V54[a]	Anz. irrel. bew. Dok. / Anz. aufg. Dok.	0,2	0,23	**0,18**[b]	**0,25**	0,18	0,22	0,17	0,28
V57	Anz. richtig eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	**0,15**	**0,3** [b]	0,24	0,21	**0,08**	**0,22**	**0,4**	**0,19**
V61	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	**0,04**	**0,09**[b]	0,06	0,07	0,04	0,03	0,07	0,1

[a] Effekt kommt in SP_B neu hinzu.
[b] Dieser Mittelwert entspricht der besseren Benutzerleistung.
[c] Stichprobengröße < 40.

Tab. E.18.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerleistung bei 4-stufiger Relevanzskala in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
M20[a]	Anz. aufg. eher irrel. Dok.	**0,93**[b]	**1,72**	1,24	1,41	0,86	1	1,61	1,82
M22[a]	Anz. aufg. irrel. Dok.	**0,52**[b]	**1,11**	0,78	0,85	0,63	0,4	0,92	1,29
M26	Anz. falsch eher irrel. bew. Dok.	**1,5**	**1,21**[b]	1,32	1,39	1,56	1,44	1,08	1,33
M31[a]	Anz. falsch eher rel. bew. eher irrel. Dok.	**0,18**[b]	**0,36**	0,33	0,21	0,25	0,11	0,41	0,31
M42[a]	Anz. richtig eher irrel. bew. Dok.	**0,23**	**0,46**[b]	0,34	0,35	0,2	0,25	0,47	0,45
M44[a]	Anz. richtig irrel. bew. Dok.	**0,36**	**0,77**[b]	0,56	0,57	0,41	0,31	0,72	0,83
B16	Durchschn. Bew. rel. Dok.	**5,59**	**6,08**[b]	5,93	5,74	5,68	5,49	6,17	5,98
V36[a]	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,05**[b]	**0,1**	0,07	0,08	0,05	0,05	0,09	0,11
V37	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,5**[b]	**0,45**	0,49	0,45	0,51	0,48	0,47	0,43
V45[c]	$\frac{\text{Anz. falsch eher rel. bew. eher irrel. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	**0,06**[b]	**0,15**	0,13	0,08	0,09	0,03	0,17	0,13
V40[a,c]	$\frac{\text{Anz. falsch eher irrel. bew. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	**0,83**	**0,69**[b]	0,76	0,77	**0,91**	**0,77**	**0,61**	**0,80**
V57[a,c]	$\frac{\text{Anz. richtig eher irrel. bew. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	**0,16** **0,18**	**0,29**[b] **0,29**[b]	0,22 0,21	0,23 0,26	**0,08** **0,06**	**0,23** **0,29**	**0,35** **0,36**	**0,23** **0,20**
V58[c]	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	0,36	0,32	**0,31**	**0,38**[b]	0,32	0,4	0,29	0,36
V61[a]	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,04**	**0,07**[b]	0,05	0,06	0,04	0,04	0,06	0,08

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.

[b] Dieser Mittelwert entspricht der besseren Benutzerleistung.

[c] Stichprobengröße < 80.

Tab. E.19.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerzufriedenheit in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F01[a]	Liefert die Suchmaschine genau die Information, die Sie benötigen?	3,32	3,23	**3,49**[b]	**3,06**	3,49	3,14	3,48	2,98
F02[a]	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	3,31	3,3	**3,51**[b]	**3,1**	3,47	3,14	3,55	3,05
F03[a]	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	3,06	2,95	**3,25**[b]	**2,75**	3,22	2,89	3,28	2,61
F04[a]	Liefert die Suchmaschine genügend Information?	3,28	3,25	**3,56**[b]	**2,97**	3,53	3,03	3,59	2,91
F05[a]	Ist die Suchmaschine präzise?	3,05	2,85	**3,28**[b]	**2,62**	3,33	2,77	3,23	2,46
F06[a]	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	3,06	2,86	**3,25**[b]	**2,67**	3,26	2,85	3,24	2,48
F07[a]	Finden Sie die Präsentation der Ergebnisse hilfreich?	3,25	3,12	**3,43**[b]	**2,94**	3,37	3,13	3,49	2,75
F08[a]	Ist die Suchmaschine benutzerfreundlich?	3,93	3,76	**4,18**[b]	**3,51**	4,11	3,75	4,24	3,28
F09[a]	Ist die Suchmaschine einfach zu bedienen?	4,43	4,49	**4,64**[b]	**4,27**	4,6	4,25	4,69	4,28

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.

Fortsetzung auf nächster Seite

[b] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.

[c] Entspricht auch der Skala SK13-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK18-M.

Tab. E.19 (Fortsetzung)

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F10[a]	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	3,69	3,69	**3,96**[b]	**3,42**	3,84	3,54	4,08	3,29
F12[a]	Ist die Suchmaschine erfolgreich?	3,36	3,39	**3,71**[b]	**3,05**	3,61	3,11	3,8	2,98
F13[a]	Sind Sie mit der Suchmaschine zufrieden?	3,34	3,25	**3,62**[b]	**2,97**	3,57	3,11	3,67	2,83
F14[a]	Es war einfach, die Aufgabe zu bearbeiten.	3,91	3,85	**4,05**[b]	**3,71**	4,1	3,71	4	3,7
F16[a]	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	3,41	3,4	**3,67**[b]	**3,14**	3,63	3,18	3,7	3,09
F17[a]	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	3,32	3,17	**3,48**[b]	**3,01**	3,49	3,14	3,46	2,87
F18[a]	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	3,39	3,11	**3,46**[b]	**3,04**	3,57	3,2	3,34	2,87
F19[a]	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**3,18**[b]	**2,89**	**3,29**[b]	**2,77**	3,34	3,01	3,24	2,53
F20[a]	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	2,73	2,73	**2,98**[b]	**2,48**	2,94	2,52	3,02	2,43
F22[a]	Ich bin mit den Suchergebnissen zufrieden.	3,35	3,23	**3,55**[b]	**3,03**	3,56	3,13	3,53	2,93
F23[a]	Ich bin mit meiner Suchleistung zufrieden.	3,4	3,32	**3,54**[b]	**3,18**	3,54	3,26	3,54	3,09
F24[a]	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	3,22	3,06	**3,4**[b]	**2,87**	3,44	2,99	3,36	2,75
F25[a]	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	3,24	3,16	**3,5**[b]	**2,89**	3,47	3,01	3,53	2,78
F26[a]	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	2,73	2,52	**3,01**[b]	**2,24**	2,97	2,48	3,05	1,99
SK01-M[a]	Genauigkeit (Mittelwert)	3,08	2,99	**3,35**[b]	**2,72**	3,36	2,79	3,34	2,64
SK02-M[a]	Inhalt (Mittelwert)	3,18	3,15	**3,44**[b]	**2,9**	3,41	2,95	3,46	2,84
SK03-M[a]	Benutzerfreundlichkeit (Mittelwert)	3,53	3,46	**3,8**[b]	**3,19**	3,74	3,31	3,85	3,06
SK04-M[a]	Suche (Mittelwert)	3,35	3,18	**3,5**[b]	**3,03**	3,52	3,17	3,48	2,88
SK07-M[ae]	Benutzerfreundlichkeit (Mittelwert)	3,24	3,11	**3,42**[b]	**2,93**	3,47	3,01	3,37	2,84
SK08-M[a]	Suche (Mittelwert)	3,3	3,14	**3,45**[b]	**2,99**	3,45	3,15	3,44	2,83
SK09-M[a]	Benutzerfreundlichkeit (Mittelwert)	3,49	3,32	**3,65**[b]	**3,16**	3,64	3,34	3,66	2,97
SK11-M[a,d]	Eigenleistung (Mittelwert)	3,02	3,06	**3,23**[b]	**2,85**	3,24	2,79	3,21	2,9
SK12-M[a]	Suchergebnis (Mittelwert)	3,2	3,14	**3,4**[b]	**2,94**	3,38	3,02	3,42	2,85
SK12-F[a]	Suchergebnis (Faktorwert)	0,08	-0,05	**0,39**[b]	**-0,36**	0,37	-0,22	0,41	-0,5
SK13-F[a]	Benutzerfreundlichkeit (Faktorwert)	0,04	-0,1	**0,38**[b]	**-0,43**	0,31	-0,23	0,44	-0,63
SK14-M[a]	Suche (Mittelwert)	3,29	3,06	**3,39**[b]	**2,95**	3,47	3,1	3,31	2,8
SK14-F[a]	Suche (Faktorwert)	0,16	-0,13	**0,36**[b]	**-0,32**	0,41	-0,08	0,3	-0,57
SK16-M[a]	Aufgabe (Mittelwert)	3,68	3,57	**3,84**[b]	**3,4**	3,87	3,48	3,81	3,32
SK16-F[a]	Aufgabe (Faktorwert)	0,03	-0,13	**0,26**[b]	**-0,35**	0,33	-0,27	0,18	-0,43
SK17-M[a]	Suche (Mittelwert)	3,2	3,08	**3,41**[b]	**2,87**	3,41	2,99	3,41	2,75
SK17-F[a]	Suche (Faktorwert)	0,13	-0,07	**0,42**[b]	**-0,37**	0,4	-0,15	0,44	-0,58
SK18-F[a]	Benutzerfreundlichkeit (Faktorwert)	0,16	-0,17	**0,32**[b]	**-0,33**	0,37	-0,05	0,27	-0,61
SK19-F[a]	Eigenleistung (Faktorwert)	0	-0,07	**0,21**[b]	**-0,28**	0,22	-0,22	0,2	-0,35
SK-A[a]	Accuracy (EUCS)	3	2,91	**3,31**[b]	**2,61**	3,3	2,7	3,31	2,51
SK-C[a]	Content (EUCS)	3,26	3,19	**3,48**[b]	**2,97**	3,43	3,08	3,53	2,85
SK-E[ca]	Ease of Use (EUCS)	4,2	4,1	**4,39**[b]	**3,91**	4,36	4,03	4,43	3,78
SK-T[a]	Timeliness (EUCS)	3,58	3,55	**3,79**[b]	**3,34**	3,75	3,4	3,82	3,28
SK-K[a]	Kriteriumsskala	3,34	3,31	**3,67**[b]	**2,98**	3,59	3,09	3,74	2,87
SK-E-88[a]	EUCS-Skala-1988	3,5	3,41	**3,68**[b]	**3,22**	3,65	3,34	3,7	3,11
SK-E-09[a]	EUCS-Skala-2009	3,26	3,2	**3,5**[b]	**2,96**	3,47	3,04	3,53	2,87
SK-E-13[a]	EUCS-Skala-2013	3,38	3,3	**3,61**[b]	**3,07**	3,6	3,16	3,62	2,97
SK-Z-13[a]	Zusatzskala-2013	3,29	3,21	**3,47**[b]	**3,02**	3,51	3,06	3,44	2,98
SK-G-13[a]	Gesamtskala-2013	3,22	3,13	**3,44**[b]	**2,9**	3,45	2,98	3,43	2,82

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.

Fortsetzung auf nächster Seite

[b] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.

[c] Entspricht auch der Skala SK13-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK18-M.

Tab. E.19 (Fortsetzung)

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
E02[a]	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	3,41	3,22	**3,6**[b]	**3,03**	3,66	3,16	3,54	2,9
E03[a]	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	3,33	3,18	**3,54**[b]	**2,97**	3,56	3,09	3,51	2,86
E04[a]	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	3,14	3,05	**3,31**[b]	**2,88**	3,31	2,97	3,31	2,79
E05[a]	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	**3,17**[b]	**2,84**	**3,32**[b]	**2,69**	3,39	2,94	3,24	2,43
		3,13[b]	**2,86**	**3,31**[b]	**2,68**	3,39	2,87	3,23	2,49
E06-M[a]	Erwartungsskala	3,28	3,13	**3,44**[b]	**2,97**	3,48	3,07	3,4	2,86

[a] Dieser Effekt wird (zum Teil) von SP_B bestätigt.
[b] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.
[c] Entspricht auch der Skala SK13-M.
[d] Entspricht auch den Skalen SK15-M und SK19-M.
[e] Entspricht auch der Skala SK18-M.

Tab. E.20.: Interaktionsmittelwerte der Varianzanalyse zur Untersuchung des Einflusses von System und Erwartung auf die Benutzerzufriedenheit in SP_B. Neben den in Tabelle 7.24 berichteten in SP_B neu hinzukommenden Effekten beinhaltet diese Tabelle auch die Gruppenmittelwerte der Effekte, die aus SP_A bestätigt werden. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	3,32	3,18	**3,53**[c]	**2,97**	3,6	3,03	3,45	2,9
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	3,33	3,26	**3,64**[c]	**2,95**	3,62	3,03	3,65	2,87
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	3,05	2,84	**3,29**[c]	**2,6**	3,27	2,82	3,3	2,38
F04	Liefert die Suchmaschine genügend Information?	3,26	3,22	**3,57**[c]	**2,91**	3,53	2,98	3,6	2,83
F05	Ist die Suchmaschine präzise?	3,02	2,83	**3,37**[c]	**2,48**	3,42	2,62	3,32	2,33
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	3,03	2,88	**3,38**[c]	**2,53**	3,42	2,63	3,33	2,42
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	3,31	2,98	**3,49**[c]	**2,8**	3,52	3,1	3,45	2,5
F08	Ist die Suchmaschine benutzerfreundlich?	3,85	3,79	**4,22**[c]	**3,42**	4,13	3,57	4,3	3,27
F09	Ist die Suchmaschine einfach zu bedienen?	4,42	4,4	**4,69**[c]	**4,13**	4,63	4,2	4,75	4,05
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	3,67	3,64	**4,02**[c]	**3,29**	4,03	3,3	4	3,28
F12	Ist die Suchmaschine erfolgreich?	3,41	3,42	**3,81**[c]	**3,02**	3,72	3,1	3,9	2,93
F13	Sind Sie mit der Suchmaschine zufrieden?	3,35	3,25	**3,68**[c]	**2,93**	3,67	3,03	3,68	2,82
F14	Es war einfach, die Aufgabe zu bearbeiten.	3,94	3,76	**4,15**[c]	**3,55**	4,28	3,6	4,02	3,5
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	3,53	3,4	**3,81**[c]	**3,12**	3,8	3,25	3,82	2,98
F17[a]	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	**3,43**[c]	**2,93**	**3,49**[c]	**2,87**	3,7	3,15	3,27	2,58
		3,43[c]	**3,04**	**3,53**[c]	**2,94**	3,7	3,15	3,35	2,72
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	3,37	3,07	**3,51**[c]	**2,93**	3,65	3,08	3,37	2,77

[a] Systemeffekt kommt in SP_B neu hinzu.
[b] Erwartungseffekt kommt in SP_B neu hinzu.
[c] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.
[d] Entspricht auch den Skalen SK15-M und SK19-M.
[e] Entspricht auch der Skala SK18-M.
[f] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.20 (Fortsetzung)

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	**3,11**[c]	**2,75**	**3,26**[c]	**2,6**	3,4	2,82	3,12	2,38
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	2,68	2,77	**3,1**[c]	**2,34**	3,03	2,32	3,17	2,37
F22	Ich bin mit den Suchergebnissen zufrieden.	3,36	3,11	**3,61**[c]	**2,86**	3,67	3,05	3,55	2,67
F23	Ich bin mit meiner Suchleistung zufrieden.	3,3	3,39	**3,68**[c]	**3,01**	3,58	3,02	3,77	3
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	3,24	2,95	**3,39**[c]	**2,8**	3,52	2,95	3,25	2,65
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	3,24	3,06	**3,56**[c]	**2,74**	3,55	2,92	3,57	2,55
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	2,75	2,43	**2,97**[c]	**2,21**	3,15	2,35	2,8	2,07
SK01-M	Genauigkeit (Mittelwert)	3,12	2,96	**3,42**[c]	**2,66**	3,48	2,76	3,36	2,56
SK02-M	Inhalt (Mittelwert)	3,21	3,1	**3,45**[c]	**2,85**	3,47	2,94	3,43	2,76
SK03-M	Benutzerfreundlichkeit (Mittelwert)	3,58	3,32	**3,85**[c]	**3,05**	3,83	3,33	3,87	2,77
SK04-M	Suche (Mittelwert)	3,34	3,11	**3,55**[c]	**2,9**	3,62	3,05	3,47	2,74
SK07-M[a,e]	Benutzerfreundlichkeit (Mittelwert)	**3,33**[c]	**2,93**	**3,41**[c]	**2,85**	3,61	3,05	3,21	2,65
		3,33[c]	**3,01**	**3,47**[c]	**2,86**	3,61	3,05	3,34	2,67
SK08-M	Suche (Mittelwert)	3,26	3,1	**3,51**[c]	**2,86**	3,55	2,98	3,47	2,73
SK09-M	Benutzerfreundlichkeit (Mittelwert)	3,47	3,24	**3,7**[c]	**3,01**	3,72	3,21	3,67	2,8
SK11-M[d]	Eigenleistung (Mittelwert)	2,99	3,11	**3,41**[c]	**2,69**	3,31	2,67	3,5	2,71
SK12-M	Suchergebnis (Mittelwert)	3,19	3,01	**3,45**[c]	**2,75**	3,48	2,89	3,41	2,61
SK12-F	Suchergebnis (Faktorwert)	0,09	-0,23	**0,44**[c]	**-0,57**	0,51	-0,33	0,36	-0,81
SK13-F	Benutzerfreundlichkeit (Faktorwert)	0,01	-0,15	**0,43**[c]	**-0,57**	0,37	-0,36	0,48	-0,78
SK14-M[a]	Suche (Mittelwert)	**3,3**[c]	**2,93**	**3,43**[c]	**2,81**	3,58	3,02	3,27	2,59
SK14-F[a]	Suche (Faktorwert)	**0,22**[c]	**-0,32**	**0,32**[c]	**-0,43**	0,56	-0,13	0,08	-0,72
		0,22[c]	**-0,3**	**0,36**[c]	**-0,44**	0,56	-0,13	0,15	-0,74
SK15-F[b]	Eigenleistung (Faktorwert)	-0,14	0,01	**0,35**[c]	**-0,47**	0,21	-0,48	0,49	-0,46
SK16-M	Aufgabe (Mittelwert)	3,73	3,56	**3,97**[c]	**3,32**	4,04	3,42	3,89	3,22
SK16-F	Aufgabe (Faktorwert)	0,15	-0,17	**0,46**[c]	**-0,48**	0,59	-0,29	0,33	-0,66
SK17-M	Suche (Mittelwert)	3,2	3,05	**3,46**[c]	**2,79**	3,48	2,92	3,44	2,66
SK17-F	Suche (Faktorwert)	0,1	-0,21	**0,39**[c]	**-0,5**	0,49	-0,28	0,29	-0,71
SK18-F[a]	Benutzerfreundlichkeit (Faktorwert)	**0,25**[c]	**-0,35**	**0,29**[c]	**-0,39**	0,55	-0,05	0,03	-0,72
SK18-F	Benutzerfreundlichkeit (Faktorwert)	**0,25**[c]	**-0,33**	**0,33**[c]	**-0,41**	0,55	-0,05	0,11	-0,76
SK19-F	Eigenleistung (Faktorwert)	-0,13	-0,07	**0,29**[c]	**-0,49**	0,25	-0,51	0,32	-0,46
SK-A	Accuracy (EUCS)	3,02	2,8	**3,32**[c]	**2,5**	3,42	2,62	3,22	2,38
SK-C	Content (EUCS)	3,24	3,08	**3,49**[c]	**2,83**	3,5	2,97	3,47	2,68
SK-E[f]	Ease of Use (EUCS)	4,13	4,04	**4,41**[c]	**3,76**	4,38	3,88	4,44	3,64
SK-T	Timeliness (EUCS)	3,6	3,56	**3,83**[c]	**3,33**	3,9	3,3	3,76	3,36
SK-K	Kriteriumsskala	3,38	3,26	**3,72**[c]	**2,92**	3,69	3,07	3,75	2,77
SK-E-88	EUCS-Skala-1988	3,45	3,3	**3,69**[c]	**3,05**	3,74	3,15	3,64	2,95
SK-E-09	EUCS-Skala-2009	3,27	3,13	**3,57**[c]	**2,83**	3,56	2,97	3,58	2,68
SK-E-13	EUCS-Skala-2013	3,4	3,25	**3,69**[c]	**2,96**	3,68	3,11	3,69	2,81
SK-Z-13	Zusatzskala-2013	3,32	3,17	**3,56**[c]	**2,93**	3,62	3,02	3,5	2,84
SK-G-13	Gesamtskala-2013	3,25	3,08	**3,49**[c]	**2,84**	3,54	2,96	3,43	2,72
E02[a]	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	**3,44**[c]	**3,15**	**3,62**[c]	**2,98**	3,78	3,1	3,45	2,85
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	3,38	3,07	**3,59**[c]	**2,86**	3,73	3,03	3,45	2,68

[a] Systemeffekt kommt in SP_B neu hinzu.

Fortsetzung auf nächster Seite

[b] Erwartungseffekt kommt in SP_B neu hinzu.

[c] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK18-M.

[f] Entspricht auch der Skala SK13-M.

Tab. E.20 (Fortsetzung)

		System		Erwartung		Interaktion			
ID	**Beschreibung**	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	3,21	2,96	**3,38**[c]	**2,79**	3,42	3	3,33	2,58
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	**3,17**[c]	**2,76**	**3,4**[c]	**2,53**	3,5	2,83	3,3	2,22
E06-M	Erwartungsskala	3,3	3,06	**3,53**[c]	**2,83**	3,61	2,99	3,44	2,67

[a] Systemeffekt kommt in SP_B neu hinzu.
[b] Erwartungseffekt kommt in SP_B neu hinzu.
[c] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit/-erwartung.
[d] Entspricht auch den Skalen SK15-M und SK19-M.
[e] Entspricht auch der Skala SK18-M.
[f] Entspricht auch der Skala SK13-M.

E.3.3. Teststatistiken der Varianzanalysen

Die in diesem Abschnitt dargestellten Tabellen enthalten weiterführende Informationen zu den im Rahmen des dritten Experiments durchgeführten Varianzanalysen. Neben der jeweiligen Stichprobengröße, geben sie Auskunft darüber, ob eine klassische (K) oder robuste (R) Varianzanalyse durchgeführt wurde oder es sich um einen eindeutigen (E) oder tendenziellen (T) Effekt handelt. In den mit *test* überschriebenen Spalten wird im Fall der klassischen Analyse der erreichte F-Wert angegeben, für robuste Analysen hingegen der entsprechende Testwert. Die mit *p* überschriebenen Spalten enthalten das zugehörige Signifikanzniveau. Die Freiheitsgrade der klassischen Analyse sind mit df abgekürzt. Im Fall der Benutzerleistung sind die Analysen der binären und 4-stufigen Relevanzskala diesmal in einer gemeinsamen Tabelle zusammenfasst.

Tab. E.21.: Teststatistik der Varianzanalyse zur Untersuchung des Einflusses von System, Erwartung und deren Wechselwirkung auf die Benutzerleistung in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**V**	**Q**	**df**	**test**	**p**	**test**	**p**	**test**	**p**
M02	Anz. aufg. Dok. (erste 10 Dok.)	116	R	E	-	1,1	0,3	0,15	0,697	9,87	**0,003**
M05	Anz. aufg. irrel. Dok.	116	R	E	-	16,41	**0,001**	1,09	0,302	0,09	0,761
M07	Anz. falsch irrel. bew. Dok.	116	R	E	-	9,74	**0,003**	0,09	0,768	0,93	0,338
M08	Anz. falsch rel. bew. Dok.	116	R	E	-	12,24	**0,001**	$1{\cdot}10^{-31}$	0,999	0,76	0,386
M14	Anz. richtig bew. Dok.	116	R	E	-	10,13	**0,003**	1,05	0,31	0,14	0,713
M15	Anz. richtig irrel. bew. Dok.	116	R	E	-	23,59	**0,001**	5,54	0,022	0,2	0,654
M20	Anz. aufg. eher irrel. Dok.	116	R	E	-	28,83	**0,001**	1,88	0,176	0,21	0,65
M22	Anz. aufg. irrel. Dok.	116	R	E	-	18,72	**0,001**	0,15	0,705	1,96	0,167
M31	Anz. falsch eher rel. bew. eher irrel. Dok.	116	R	E	-	6,38	**0,014**	0,71	0,403	0,01	0,905
M26	Anz. falsch eher irrel. bew. Dok.	116	R	T	-	5,15	**0,027**	0,08	0,783	1,35	0,25
M42	Anz. richtig eher irrel. bew. Dok.	116	R	E	-	11,65	**0,002**	0,3	0,585	0,11	0,743
M44	Anz. richtig irrel. bew. Dok.	116	R	E	-	10,29	**0,003**	0,01	0,932	0,61	0,439
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	96	K	E	1/92	6,59	**0,012**	0,92	0,339	1,03	0,313
B04	Durchschn. Bew. rel. Dok.	116	K	E	1/112	8,41	**0,005**	0,29	0,591	0,11	0,736
B16	Durchschn. Bew. rel. Dok.	116	K	E	1/112	7,96	**0,006**	1,15	0,286	$4{\cdot}10^{-4}$	0,983
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	80	R	T	-	7,72	**0,009**	1,11	0,301	0,91	0,347
Z02-log	Durchschn. Betrachtungsz. falsch bew. Dok.	80	R	E	-	11,61	**0,002**	2,42	0,13	1	0,324

Fortsetzung auf nächster Seite

Tab. E.21 (Fortsetzung)

ID	Beschreibung	n	V	Q	df	System test	System p	Erwartung test	Erwartung p	Interaktion test	Interaktion p
Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	88	R	E	-	8,39	**0,006**	0,21	0,653	0,57	0,455
Z05-log	Durchschn. Betrachtungsz. irrel. bew. Dok.	88	R	E	-	8,96	**0,005**	0,45	0,507	0,13	0,725
V01	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	116	K	E	1/112	69,49	$\mathbf{2{\cdot}10^{-13}}$	1,59	0,21	0,01	0,935
V02	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	116	K	E	1/112	65,4	$\mathbf{1{\cdot}10^{-12}}$	2,05	0,155	$1{\cdot}10^{-3}$	0,978
V05	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	116	R	E	-	15,18	**0,001**	1,54	0,22	0,22	0,639
V06	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	88	K	E	1/84	37,12	$\mathbf{3{\cdot}10^{-8}}$	2,39	0,126	0,18	0,671
V08	Anz. falsch rel. bew. Dok. / Anz. aufg. Dok.	116	R	E	-	12,66	**0,001**	1,35	0,25	0,12	0,734
V09	Anz. falsch rel. bew. Dok. / Anz. rel. bew. Dok.	108	R	E	-	13,48	**0,001**	0,74	0,394	$1{\cdot}10^{-4}$	0,995
V10	Anz. falsch rel. bew. Dok. / Anz. richtig rel. bew. Dok.	108	R	E	-	13,42	**0,001**	0,87	0,355	$2{\cdot}10^{-3}$	0,962
V13	Anz. richtig bew. Dok. / Anz. aufg. Dok.	116	R	E	-	17,4	**0,001**	1,7	0,198	0,68	0,413
V14	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	116	K	E	1/112	43,79	$\mathbf{1{\cdot}10^{-9}}$	4,42	**0,038**	0,2	0,657
						35,76	$\mathbf{3{\cdot}10^{-8}}$	5,58	**0,02**	0,01	0,934
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	88	K	E	1/84	40,29	$\mathbf{1{\cdot}10^{-8}}$	5,04	0,027	0,12	0,729
V31/BP	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.	108	R	E	-	14,13	**0,001**	0,57	0,456	0,69	0,41
V36	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	116	R	E	-	15,19	**0,001**	0,91	0,343	1,11	0,297
V37	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	116	K	E	1/112	6,58	**0,012**	3,44	0,066	0,08	0,784
V45	Anz. falsch eher rel. bew. eher irrel. Dok. / Anz. eher rel. bew. Dok.	68	R	E	-	12,16	**0,002**	4,17	0,05	0,41	0,529
V40	Anz. falsch eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	52	R	E	-	10,7	**0,003**	0,37	0,548	14,58	**0,001**
V57	Anz. richtig eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	52	R	E	-	6,77	**0,015**	0,02	0,889	7,25	**0,012**
						4,47	**0,044**	0,5	0,486	14,04	**0,001**
V58	Anz. richtig eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	68	R	T	-	0,26	0,616	5,41	**0,026**	0,11	0,74
V61	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	116	R	E	-	9,7	**0,003**	1,05	0,309	0,73	0,397

Tab. E.22.: Teststatistik der Varianzanalyse zur Untersuchung des Einflusses von System, Erwartung und deren Wechselwirkung auf die Benutzerleistung in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

ID	Beschreibung	n	V	Q	df	System test	System p	Erwartung test	Erwartung p	Interaktion test	Interaktion p
M02	Anz. aufg. Dok. (erste 10 Dok.)	80	R	E		0,45	0,509	1,9	0,176	8,89	**0,005**
M05	Anz. aufg. irrel. Dok.	80	R	E	-	19,99	**0,001**	0,3	0,592	0,21	0,65
M07	Anz. falsch irrel. bew. Dok.	80	R	E	-	8,44	**0,006**	0,36	0,553	0,77	0,386

Fortsetzung auf nächster Seite

Tab. E.22 (Fortsetzung)

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**V**	**Q**	**df**	**test**	**p**	**test**	**p**	**test**	**p**
M08	Anz. falsch rel. bew. Dok.	80	R	T	-	9,4	**0,004**	3,38	0,074	0,04	0,84
M14	Anz. richtig bew. Dok.	80	R	E	-	8,92	**0,005**	0,02	0,892	0,43	0,519
M15	Anz. richtig irrel. bew. Dok.	80	R	E	-	18,57	**0,001**	0,86	0,363	0,03	0,872
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	80	R	T	-	1,58	0,217	1,91	0,175	5,12	**0,03**
M19	Anz. richtig rel. bew. Dok. (letzte Suche)	80	R	T	-	0,64	0,429	5,29	**0,028**	0,05	0,818
M20	Anz. aufg. eher irrel. Dok.	80	R	E	-	20,54	**0,001**	0,55	0,462	2,02	0,164
M22	Anz. aufg. irrel. Dok.	80	R	E	-	16,57	**0,001**	0,05	0,826	0,05	0,826
M27	Anz. falsch eher irrel. bew. eher rel. Dok.	80	R	E	-	5,27	0,031	0,74	0,399	13,93	**0,001**
M31	Anz. falsch eher rel. bew. eher irrel. Dok.	80	R	E	-	8,78	**0,006**	0,98	0,33	0,02	0,889
M42	Anz. richtig eher irrel. bew. Dok.	80	R	E	-	12,45	**0,002**	1,15	0,291	1,91	0,176
M44	Anz. richtig irrel. bew. Dok.	80	R	E	-	12,63	**0,001**	0,44	0,51	3,16	0,084
B04	Durchschn. Bew. rel. Dok.	80	K	E	1/76	1,71	0,195	7,49	**0,008**	1,24	0,269
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	68	K	E	1/64	3,38	0,071	9,37	**0,003**	0,99	0,323
B16	Durchschn. Bew. rel. Dok.	76	K	E	1/72	2,1	0,151	8,16	**0,006**	0,72	0,397
B18	Durchschn. Bew. rel. Dok. (letzte Suche)	56	R	T	-	0,86	0,361	7,19	**0,012**	0,39	0,536
Z01	Durchschn. Betrachtungsz. aller Dok.	80	R	E	-	13,79	**0,001**	2,66	0,113	$3{\cdot}10^{-5}$	0,996
Z01-log	Durchschn. Betrachtungsz. aller Dok.	80	R	E	-	13,17	**0,001**	1,2	0,28	0,17	0,679
Z02-log	Durchschn. Betrachtungsz. falsch bew. Dok.	52	R	T	-	11,74	**0,002**	0,74	0,398	0,45	0,51
Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	56	R	E	-	15,19 13,66	**0,001** **0,001**	0,05 $2{\cdot}10^{-3}$	0,819 0,965	5,69 6,39	**0,023** **0,017**
Z05-log	Durchschn. Betrachtungsz. irrel. bew. Dok.	56	R	E	-	18,68	**0,001**	0,1	0,756	3,67	0,064
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	76	R	E	-	7,99	**0,008**	0,79	0,379	0,6	0,443
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	72	R	T	-	6,71	**0,014**	2,28	0,139	0,4	0,529
Z15	Durchschn. Betrachtungsz. eher rel. Dok.	52	R	E	-	17,75	**0,001**	$3{\cdot}10^{-3}$	0,958	0,08	0,785
V01	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	80	K	E	1/76	50,51	$\mathbf{1{\cdot}10^{-9}}$	0,3	0,585	0,07	0,796
V02	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	80	K	E	1/76	54,5	$\mathbf{2{\cdot}10^{-10}}$	0,05	0,818	0,34	0,561
V05	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	80	K	E	1/76	14,49	$\mathbf{3{\cdot}10^{-4}}$	3,95	0,05	1,9	0,173
V06	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	56	K	E	1/52	28,22	$\mathbf{2{\cdot}10^{-6}}$	0,64	0,428	0,84	0,365
V08	Anz. falsch rel. bew. Dok. / Anz. aufg. Dok.	80	R	E	-	8,68 5,79	**0,006** **0,021**	4,22 5,66	**0,046** **0,022**	0,36 0,91	0,551 0,345
V09	Anz. falsch rel. bew. Dok. / Anz. rel. bew. Dok.	72	R	E	-	8,36	**0,006**	2,13	0,152	1,26	0,268
V10	Anz. falsch rel. bew. Dok. / Anz. richtig rel. bew. Dok.	72	R	E	-	7,2	**0,011**	0,55	0,463	0,62	0,435
V11	Anz. irrel. bew. Dok. / Anz. aufg. Dok.	80	K	E	1/76	$1{\cdot}10^{-4}$	0,993	5,82	**0,018**	1,49	0,227
V12	Anz. rel. bew. Dok. / Anz. aufg. Dok.	80	K	T	1/76	0,07	0,787	6,51	**0,013**	1,84	0,179
V13	Anz. richtig bew. Dok. / Anz. aufg. Dok.	80	K	E	1/76	13,94	$\mathbf{4{\cdot}10^{-4}}$	0,79	0,375	0,14	0,714
V14	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	80	K	E	1/76	43,44	$\mathbf{1{\cdot}10^{-8}}$	2,73	0,103	0,09	0,768
V16	Anz. richtig irrel. bew. Dok. / Anz. falsch irrel. bew. Dok.	40	R	E	-	9,49	**0,008**	1,01	0,334	3,02	0,106

Fortsetzung auf nächster Seite

Tab. E.22 (Fortsetzung)

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**V**	**Q**	**df**	**test**	**p**	**test**	**p**	**test**	**p**
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	56	K	E	1/52	24,31	$\mathbf{1{\cdot}10^{-5}}$	0,2	0,653	1,89	0,175
V29	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	72	K	T	1/68	9,47	**0,003**	1,4	0,241	1,6	0,21
V31/BP	Anz. richtig rel. bew. Dok. / Anz. rel. bew. Dok.	72	R	E	-	6,77	**0,013**	1,2	0,28	0,51	0,478
V33	Anz. richtig rel. bew. Dok. / Anz. zurückgeg. rel. Dok.	72	R	T	-	5,89	**0,021**	2,81	0,103	0,04	0,835
V36	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	80	R	E	-	15,38	**0,001**	0,29	0,592	0,63	0,433
V40	Anz. falsch eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	36	R	E	-	7,05	**0,015**	1,17	0,293	14,06	**0,002**
V41	Anz. falsch eher irrel. bew. eher rel. Dok. / Anz. eher irrel. bew. Dok.	36	R	T	-	6,39	**0,023**	4,64	0,048	1,48	0,243
V54	Anz. irrel. bew. Dok. / Anz. aufg. Dok.	80	K	T	1/76	0,81	0,371	7,18	**0,009**	1,68	0,198
V57	Anz. richtig eher irrel. bew. Dok. / Anz. eher irrel. bew. Dok.	36	R	T	-	7,12	0,015	1,41	0,25	14,13	**0,002**
V61	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	80	R	E	-	17,41	**0,001**	0,42	0,521	1,25	0,27

Tab. E.23.: Teststatistik der Varianzanalyse zur Untersuchung des Einflusses von System, Erwartung und deren Wechselwirkung auf die Benutzerzufriedenheit in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**V**	**Q**	**df**	**test**	**p**	**test**	**p**	**test**	**p**
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	116	R	E	-	1,38	0,247	15,51	**0,001**	$5{\cdot}10^{-3}$	0,946
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	116	R	E	-	0,52	0,475	12,03	**0,001**	0,04	0,846
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	116	K	E	1/112	0,68	0,41	14,35	$\mathbf{2{\cdot}10^{-4}}$	1,59	0,209
F04	Liefert die Suchmaschine genügend Information?	116	K	E	1/112	0,05	0,827	13,82	$\mathbf{3{\cdot}10^{-4}}$	0,34	0,561
F05	Ist die Suchmaschine präzise?	116	K	E	1/112	2,04	0,156	21,19	$\mathbf{1{\cdot}10^{-5}}$	0,51	0,477
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	116	R	E	-	1,67	0,201	18,89	**0,001**	1,67	0,201
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	116	K	E	1/112	0,74	0,392	11,26	**0,001**	2,95	0,089
F08	Ist die Suchmaschine benutzerfreundlich?	116	R	E	-	3,39	0,071	17,51	**0,001**	4,26	0,044
F09	Ist die Suchmaschine einfach zu bedienen?	116	R	E	-	0,04	0,85	8,93	**0,004**	0,2	0,658
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	116	R	E	-	0,14	0,71	12,84	**0,001**	2,41	0,126
F12	Ist die Suchmaschine erfolgreich?	116	K	E	1/112	0,04	0,837	22,42	$\mathbf{1{\cdot}10^{-5}}$	1,43	0,235
F13	Sind Sie mit der Suchmaschine zufrieden?	116	K	E	1/112	0,46	0,498	20,43	$\mathbf{2{\cdot}10^{-5}}$	1,74	0,19
F14	Es war einfach, die Aufgabe zu bearbeiten.	116	K	E	1/112	0,23	0,635	8,15	**0,005**	0,14	0,704
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	116	R	E	-	0,38	0,538	16	**0,001**	0,38	0,538
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	116	R	T	-	1,36	0,248	9,68	**0,003**	2,13	0,15

Fortsetzung auf nächster Seite

Tab. E.23 (Fortsetzung)

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**V**	**Q**	**df**	**test**	**p**	**test**	**p**	**test**	**p**
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	116	R	E	-	5,82	0,019	11,94	**0,001**	0,04	0,846
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	116	R	E	-	6,9	**0,011**	11,87	**0,001**	2,87	0,095
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	116	K	E	1/112	$1{\cdot}10^{-3}$	0,973	9,12	**0,003**	0,26	0,612
F22	Ich bin mit den Suchergebnissen zufrieden.	116	R	E	-	1,32	0,255	9,3	**0,004**	0,2	0,655
F23	Ich bin mit meiner Suchleistung zufrieden.	116	K	E	1/112	0,39	0,535	6,82	**0,01**	0,39	0,535
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	116	K	E	1/112	1,27	0,262	13,7	$\mathbf{3{\cdot}10^{-4}}$	0,32	0,574
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	116	K	E	1/112	0,32	0,574	15,62	$\mathbf{1{\cdot}10^{-4}}$	0,89	0,349
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	116	R	E	-	1,58	0,213	18,47	**0,001**	2,31	0,134
SK01-M	Genauigkeit	116	K	E	1/112	0,48	0,491	26,46	$\mathbf{1{\cdot}10^{-6}}$	0,26	0,613
SK02-M	Inhalt	116	R	E	-	0,39	0,535	18,06	**0,001**	0,39	0,535
SK03-M	Benutzerfreundlichkeit	116	K	E	1/112	0,33	0,567	23,96	$\mathbf{3{\cdot}10^{-6}}$	2,1	0,15
SK04-M	Suche	116	R	E	-	2,99	0,089	12,91	**0,001**	0,83	0,367
SK07-M[b]	Benutzerfreundlichkeit	116	K	E	1/112	0,88	0,35	13,03	$\mathbf{1{\cdot}10^{-3}}$	0,07	0,787
SK08-M	Suche	116	R	E	-	3,33	0,073	12,37	**0,001**	1,7	0,197
SK09-M	Benutzerfreundlichkeit	116	R	E	-	5,77	0,02	14,42	**0,001**	4,38	0,041
SK11-M[a]	Eigenleistung	116	K	E	1/112	0,08	0,777	7,16	**0,009**	0,24	0,628
SK12-M	Suchergebnis	116	K	E	1/112	0,31	0,577	14,8	$\mathbf{2{\cdot}10^{-4}}$	0,73	0,394
SK12-F	Suchergebnis	116	K	E	1/112	0,47	0,496	18,88	$\mathbf{3{\cdot}10^{-5}}$	0,81	0,37
SK13-F	Benutzerfreundlichkeit	116	R	E	-	1,45	0,234	15,85	**0,001**	2,68	0,107
SK14-M	Suche	116	K	E	1/112	3,31	0,072	11,75	**0,001**	0,29	0,593
SK14-F	Suche	116	R	E	-	2,9	0,094	12,52	**0,001**	1,42	0,238
SK16-M	Aufgabe	116	R	E	-	1,88	0,176	14,3	**0,001**	0,03	0,868
SK16-F	Aufgabe	116	R	E	-	1,73	0,193	14,28	**0,001**	0,19	0,661
SK17-M	Suche	116	K	E	1/112	1,01	0,317	21,22	$\mathbf{1{\cdot}10^{-5}}$	1,1	0,297
SK17-F	Suche	116	K	E	1/112	1,22	0,272	20,1	$\mathbf{2{\cdot}10^{-5}}$	1,77	0,186
SK18-F	Benutzerfreundlichkeit	116	K	E	1/112	4,03	0,047	15,57	$\mathbf{1{\cdot}10^{-4}}$	1,93	0,168
SK19-F	Eigenleistung	116	K	T	1/112	0,15	0,696	7,72	**0,006**	0,09	0,77
SK-A	Accuracy (EUCS)	116	K	E	1/112	0,42	0,52	25,6	$\mathbf{2{\cdot}10^{-6}}$	0,53	0,468
SK-C	Content (EUCS)	116	R	E	-	1,17	0,286	14,93	**0,001**	0,8	0,377
SK-E[c]	Ease of Use (EUCS)	116	R	E	-	1,49	0,228	12,87	**0,001**	1,85	0,179
SK-T	Timeliness (EUCS)	116	R	E	-	0,43	0,514	14,18	**0,001**	1	0,321
SK-K	Kriteriumsskala	116	K	E	1/112	0,07	0,796	26,49	$\mathbf{1{\cdot}10^{-6}}$	1,92	0,169
SK-E-88	EUCS-Skala-1988	116	K	E	1/112	0,7	0,404	16,92	$\mathbf{1{\cdot}10^{-4}}$	1,53	0,218
SK-E-09	EUCS-Skala-2009	116	K	E	1/112	0,28	0,595	21,92	$\mathbf{1{\cdot}10^{-5}}$	0,96	0,33
SK-E-13	EUCS-Skala-2013	116	K	E	1/112	0,61	0,437	23,91	$\mathbf{3{\cdot}10^{-6}}$	0,86	0,355
SK-Z-13	Zusatzskala-2013	116	K	E	1/112	0,47	0,496	16,43	$\mathbf{1{\cdot}10^{-4}}$	0,01	0,924
SK-G-13	Gesamtskala-2013	116	K	E	1/112	0,56	0,457	23,25	$\mathbf{1{\cdot}10^{-5}}$	0,38	0,538
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	116	R	E	-	4,72	0,034	14,01	**0,001**	0,71	0,402
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	116	R	E	-	1,54	0,219	15,72	**0,001**	1,01	0,319

Fortsetzung auf nächster Seite

Tab. E.23 (Fortsetzung)

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**V**	**Q**	**df**	**test**	**p**	**test**	**p**	**test**	**p**
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	116	R	E	-	2,52	0,118	11,42	**0,002**	1,07	0,305
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	116	R	E	-	7,85	**0,007**	18,38	**0,001**	1,73	0,193
			K		1/112	3,47	**0,065**	18,67	$\mathbf{3{\cdot}10^{-5}}$	0,57	0,453
E06-M	Erwartungsskala	116	R	E	-	2,93	0,092	12,84	**0,001**	0,47	0,494

[a] Entspricht auch den Skalen SK15-M und SK19-M.
[b] Entspricht auch der Skala SK18-M.
[c] Entspricht auch der Skala SK13-M.

Tab. E.24.: Teststatistik der Varianzanalyse zur Untersuchung des Einflusses von System, Erwartung und deren Wechselwirkung auf die Benutzerzufriedenheit in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**V**	**Q**	**df**	**test**	**p**	**test**	**p**	**test**	**p**
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	80	R	E	-	1,2	0,283	16,65	**0,001**	0,01	0,922
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	80	R	E	-	0,82	0,373	18,84	**0,001**	0,03	0,858
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	80	K	E	1/76	1,77	0,187	20,71	$\mathbf{2{\cdot}10^{-5}}$	2,41	0,124
F04	Liefert die Suchmaschine genügend Information?	80	R	E	-	0,47	0,496	12,94	**0,001**	0,99	0,328
F05	Ist die Suchmaschine präzise?	80	K	E	1/76	1,46	0,231	31,5	$\mathbf{3{\cdot}10^{-7}}$	0,33	0,566
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	80	K	E	1/76	0,89	0,348	28,62	$\mathbf{1{\cdot}10^{-6}}$	0,18	0,676
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	80	K	E	1/76	4,17	0,045	17,52	$\mathbf{1{\cdot}10^{-4}}$	2,67	0,107
F08	Ist die Suchmaschine benutzerfreundlich?	80	R	E	-	1,5	0,23	17,04	**0,001**	2,34	0,135
F09	Ist die Suchmaschine einfach zu bedienen?	80	R	E	-	0,07	0,795	16,98	**0,001**	1,3	0,263
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	80	K	E	1/76	0,02	0,893	15,43	$\mathbf{2{\cdot}10^{-4}}$	$2{\cdot}10^{-3}$	0,964
F12	Ist die Suchmaschine erfolgreich?	80	R	E	-	0,22	0,643	24,56	**0,001**	3,16	0,083
F13	Sind Sie mit der Suchmaschine zufrieden?	80	K	E	1/76	0,37	0,544	20,93	$\mathbf{2{\cdot}10^{-5}}$	0,51	0,479
F14	Es war einfach, die Aufgabe zu bearbeiten.	80	K	E	1/76	1,54	0,218	16,54	$\mathbf{1{\cdot}10^{-4}}$	0,32	0,574
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	80	K	E	1/76	0,73	0,397	22,23	$\mathbf{1{\cdot}10^{-5}}$	0,93	0,337
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	80	R	E	-	7,41	**0,01**	8,85	**0,005**	1,16	0,289
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	80	K	E	1/76	4,26	**0,042**	9,73	**0,003**	0,05	0,827
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	80	K	E	1/76	4,51	0,037	17,07	$\mathbf{1{\cdot}10^{-4}}$	0,01	0,906
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	80	R	E	-	10,77	**0,003**	15,78	**0,001**	1,07	0,307
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	80	K	E	1/76	0,28	0,601	18,93	$\mathbf{4{\cdot}10^{-5}}$	0,06	0,812
F22	Ich bin mit den Suchergebnissen zufrieden.	80	K	E	1/76	2,09	0,152	18,81	$\mathbf{4{\cdot}10^{-5}}$	0,59	0,443
F23	Ich bin mit meiner Suchleistung zufrieden.	80	K	E	1/76	0,28	0,6	17,73	$\mathbf{1{\cdot}10^{-4}}$	0,4	0,53

[a] Entspricht auch den Skalen SK15-M und SK19-M.
[b] Entspricht auch der Skala SK18-M.
[c] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.24 (Fortsetzung)

						System		Erwartung		Interaktion	
ID	**Beschreibung**	**n**	**V**	**Q**	**df**	**test**	**p**	**test**	**p**	**test**	**p**
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	80	K	E	1/76	2,58	0,113	10,92	**0,001**	0,01	0,925
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	80	K	E	1/76	1,07	0,305	23,69	$\mathbf{1{\cdot}10^{-5}}$	1,28	0,262
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	80	K	E	1/76	2,83	0,096	16,62	$\mathbf{1{\cdot}10^{-4}}$	0,03	0,86
SK01-M	Genauigkeit	80	K	E	1/76	1,33	0,252	29,76	$\mathbf{1{\cdot}10^{-6}}$	0,1	0,751
SK02-M	Inhalt	80	R	E	-	0,83	0,369	13,3	**0,001**	0,47	0,499
SK03-M	Benutzerfreundlichkeit	80	K	E	1/76	3,32	0,072	31,16	$\mathbf{4{\cdot}10^{-7}}$	4,47	0,038
SK04-M	Suche	80	K	E	1/76	3,13	0,081	24,7	$\mathbf{4{\cdot}10^{-6}}$	0,38	0,538
SK07-M[b]	Benutzerfreundlichkeit	80	K	E	1/76	6,42	**0,013**	12,5	**0,001**	0	1
						4,11	**0,046**	14,8	$\mathbf{2{\cdot}10^{-4}}$	0,13	0,717
SK08-M	Suche	80	K	E	1/76	1,64	0,204	26,53	$\mathbf{2{\cdot}10^{-6}}$	0,4	0,53
SK09-M	Benutzerfreundlichkeit	80	K	E	1/76	2,74	0,102	24,66	$\mathbf{4{\cdot}10^{-6}}$	1,58	0,213
SK11-M[a]	Eigenleistung	80	K	E	1/76	0,6	0,439	22,81	$\mathbf{1{\cdot}10^{-5}}$	0,25	0,619
SK12-M	Suchergebnis	80	K	E	1/76	1,69	0,198	26,42	$\mathbf{2{\cdot}10^{-6}}$	0,6	0,443
SK12-F	Suchergebnis	80	K	E	1/76	2,69	0,105	27,71	$\mathbf{1{\cdot}10^{-6}}$	0,78	0,381
SK13-F	Benutzerfreundlichkeit	80	R	E		1,56	0,22	22,84	**0,001**	3,74	0,06
SK14-M	Suche	80	R	E	-	6,65	**0,015**	14,02	**0,001**	1,05	0,312
SK14-F	Suche	80	R	E	-	8,95	**0,005**	10,61	**0,003**	0,84	0,365
						8,03	**0,008**	12,96	**0,001**	1,54	0,224
SK15-F	Eigenleistung	80	K	E	1/76	0,59	0,446	17,88	$\mathbf{1{\cdot}10^{-4}}$	0,42	0,521
SK16-M	Aufgabe	80	K	E	1/76	1,79	0,185	23,21	$\mathbf{1{\cdot}10^{-5}}$	0,05	0,828
SK16-F	Aufgabe	80	K	E	1/76	2,53	0,116	22,49	$\mathbf{1{\cdot}10^{-5}}$	0,07	0,796
SK17-M	Suche	80	K	E	1/76	1,55	0,217	28,63	$\mathbf{1{\cdot}10^{-6}}$	0,75	0,39
SK17-F	Suche	80	R	E	-	3,84	0,059	13,86	**0,001**	0,68	0,418
SK18-F	Benutzerfreundlichkeit	80	K	E	1/76	7,38	**0,008**	9,5	**0,003**	0,12	0,726
						7,29	**0,009**	11,8	**0,001**	0,4	0,528
SK19-F	Eigenleistung	80	K	E	1/76	0,08	0,775	14,4	$\mathbf{3{\cdot}10^{-4}}$	0,01	0,939
SK-A	Accuracy (EUCS)	80	K	E	1/76	2,13	0,149	28,06	$\mathbf{1{\cdot}10^{-6}}$	0,03	0,872
SK-C	Content (EUCS)	80	R	E	-	1,85	0,185	17	**0,001**	0,71	0,406
SK-E[c]	Ease of Use (EUCS)	80	R	E	-	2,07	0,158	21,51	**0,001**	2,39	0,13
SK-T	Timeliness (EUCS)	80	R	E	-	0,2	0,654	10,04	**0,004**	0,06	0,803
SK-K	Kriteriumsskala	80	K	E	1/76	0,6	0,441	28,27	$\mathbf{1{\cdot}10^{-6}}$	1,35	0,249
SK-E-88	EUCS-Skala-1988	80	K	E	1/76	1,55	0,217	27,73	$\mathbf{1{\cdot}10^{-6}}$	0,16	0,693
SK-E-09	EUCS-Skala-2009	80	K	E	1/76	1,17	0,284	32,9	$\mathbf{2{\cdot}10^{-7}}$	1,38	0,243
SK-E-13	EUCS-Skala-2013	80	K	E	1/76	1,34	0,25	33,21	$\mathbf{2{\cdot}10^{-7}}$	1,45	0,233
SK-Z-13	Zusatzskala-2013	80	K	E	1/76	1,54	0,218	26,19	$\mathbf{2{\cdot}10^{-6}}$	0,06	0,813
SK-G-13	Gesamtskala-2013	80	K	E	1/76	1,96	0,166	28,44	$\mathbf{1{\cdot}10^{-6}}$	0,27	0,605
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	80	R	E	-	8,04	**0,007**	15,47	**0,001**	0,01	0,928
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	80	R	E	-	5,63	0,023	20,25	**0,001**	1,51	0,227
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	80	R	E	-	5,23	0,028	14,52	**0,001**	2,32	0,135
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	80	K	E	1/76	6,62	**0,012**	30,42	$\mathbf{1{\cdot}10^{-6}}$	1,72	0,193
E06-M	Erwartungsskala	80	K	E	1/76	3,11	0,082	24,59	$\mathbf{4{\cdot}10^{-6}}$	0,29	0,592

[a] Entspricht auch den Skalen SK15-M und SK19-M.

[b] Entspricht auch der Skala SK18-M.

[c] Entspricht auch der Skala SK13-M.

E.4. Weitere Ergebnisse der dynamischen Analyse des Benutzerverhaltens

In diesem Abschnitt sind weitere Tabellen in Bezug auf die Analyse der dynamischen Entwicklung des Benutzerverhaltens aufgeführt. Zur besseren Einordnung der dargestellten Ergebnisse ist der Abschnitt erneut in vier Unterabschnitte gegliedert. Abschnitt E.4.1 enthält zunächst wiederum eine Übersicht über diejenigen Variablen die in keiner der fünf Stichproben einen signifikanten Effekt der untersuchten Einflussgrößen zeigen. In Abschnitt E.4.2 sind all diejenigen Gruppenmittelwerte zur Verfügung gestellt, die aus Platzgründen nicht im Hauptteil der Arbeit dargestellt wurden. Bevor in Abschnitt E.4.4 erneut die Teststatistiken der dynamischen Analyse wiedergegeben werden, umfasst Abschnitt E.4.3 die Mittelwerte der im Rahmen des Hauptteils der Arbeit besprochenen dynamischen Interaktionen.

E.4.1. Variablen ohne signifikante Unterschiede

Tabelle E.25 stellt eine Übersicht derjenigen Variablen bereit, die in keiner der fünf Stichproben einen signifikanten Effekt der untersuchten Einflussgrößen zeigen. Das Fehlen eines Effekts wird aufgeschlüsselt nach den beiden Datenqualitätsstufen SP_A und SP_B. Darüber hinaus gibt die Tabelle erneut die Stichprobengröße sowie die Art der Varianzanalyse (klassisch vs. robust) an.

Tab. E.25.: Übersicht über Variablen ohne signifikante Unterschiede. Minuszeichen (-) kennzeichnen Variablen für die aufgrund zu geringer Fallzahlen keine Analyse durchgeführt werden kann.

		SP_A		SP_B	
ID	**Beschreibung**	**n**	**V**	**n**	**V**
M03	Anz. aufg. Dok. (erste Suche)	72	R	-	-
M12	Anz. rel. bew. Dok. (erste Suche)	72	R	-	-
M13	Anz. rel. bew. Dok. (letzte Suche)	116	R	-	-
M18	Anz. richtig rel. bew. Dok. (erste Suche)	116	R	80	R
M25	Anz. eher rel. bew. Dok.	116	R	80	R
M29	Anz. falsch eher irrel. bew. rel. Dok.	116	R	-	-
M34	Anz. falsch irrel. bew. Dok. (4-st.)	72	R	-	-
M35	Anz. falsch rel. bew. Dok. (4-st.)	72	R	-	-
M36	Anz. irrel. bew. Dok. (4-st.)	72	R	-	-
M33	Anz. falsch eher rel. bew. rel. Dok.	72	R	-	-
M43	Anz. richtig eher rel. bew. Dok.	72	R	-	-
M46	Anz. richtig rel. bew. Dok. (erste 10 Dok.) (4-st.)	116	R	80	R
M47	Anz. richtig rel. bew. Dok. (erste Suche) (4-st.)	72	R	-	-
B05	Durchschn. Bew. rel. Dok. (erste Suche)	72	R	48	R
B16	Durchschn. Bew. rel. Dok. (4-st.)	-	-	76	R
B17	Durchschn. Bew. rel. Dok. (erste Suche) (4-st.)	60	R	-	-
S02	Erste betr. Rankingpos.	-	-	80	R
S03	Letzte betr. Rankingpos.	116	R	-	-
S04	Suchdauer	-	-	80	R
S06	Zeit zum ersten richtig rel. bew. Dok. (4-st.)	56	R	-	-
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	-	-	80	R
V22	$\frac{\text{Anz. richtig rel. bew. Dok. (erste Suche)}}{\text{Anz. rel. Dok. im Korpus}}$	48	R	-	-
V24	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. aufg. Dok. (letzte Suche)}}$	48	K/R	-	-
V26	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. rel. Dok. im Korpus}}$	80	R	52	R
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	-	-	72	R

Fortsetzung auf nächster Seite

Tab. E.25 (Fortsetzung)

		SP_A		SP_B	
ID	**Beschreibung**	**n**	**V**	**n**	**V**
V35	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	72	R	-	-
V48	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	R	-	-
V51	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	R	-	-
V52	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$ (4-st.)	80	R	-	-
V55	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	116	R	80	R
V59	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$	72	R	-	-
V60	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	72	R	-	-
Z28	Durchschn. Betrachtungsz. richtig rel. bew. Dok. (4-st.)	56	R	-	-

E.4.2. Gruppenmittelwerte der dynamischen Analyse

Auch im Kontext der dynamischen Analyse enthalten die im Hauptteil der Arbeit angegebenen Tabellen ausgewählte Ergebnisse der im Rahmen der Auswertung durchgeführten Analysen. So werden im Zusammenhang mit der Benutzerleistung bspw. lediglich signifikante Positionseffekte der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung in SP_A berichtet, da in SP_B keine neuen Effekte der Aufgabenposition hinzukommen. Darüber hinaus werden in den Tabellen im Hauptteil der Arbeit ausschließlich Effekte ausgewiesen, die einen signifikanten Haupteffekt der Aufgabenposition aufweisen. Ergänzend zu den bereits im Hauptteil der Arbeit berichteten Ergebnissen stellen die im Folgenden aufgeführten Tabellen daher, neben den signifikanten Positionseffekten der Benutzerleistung in SP_B, auch die signifikante Gruppenmittelwerte aus Analysen zur Verfügung, die keinen signifikanten Haupteffekt der Aufgabenposition aufweisen. Im Zusammenhang mit der Benutzerzufriedenheit ergibt sich nur für das Zufriedenheitsitem F09 in SP_A ein signifikanter Positionseffekt, weshalb die Ergebnisse des Posthoc-Tests in diesem Fall in Tabelle E.29 mit ausgewiesen werden.

Tab. E.26.: Signifikante Positionseffekte der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. Größer-/Kleinerzeichen ($\gtrless$) zwischen den Positionsmittelwerten (P_i) markieren signifikante Mittelwertsunterschiede. Dabei bezieht sich die letzte Spalte auf den Vergleich zwischen P_3 und P_1.

		System		**Erwartung**		**Position**			
ID	**Beschreibung**	$\mathbf{S_G}$	$\mathbf{S_S}$	$\mathbf{E_H}$	$\mathbf{E_N}$	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$	
M01[b]	Anz. aufg. Dok.	8,48	9,61	9,62	8,47	**7,69**	**< 9,82**[a]	**9,62**	**>**
M07[c]	Anz. falsch irrel. bew. Dok.	**2,55**	**1,67**[a]	2,05	2,17	**1,77**[a]	**2,71**	**1,84**	
M07[b]	Anz. falsch irrel. bew. Dok.	**2,55**	**1,68**[a]	2,07	2,17	**1,75**[a]	**< 2,76 >**	**1,84**	
M10[b]	Anz. rel. bew. Dok.	4,82	5,56	5,63	4,75	**4,51**	**< 5,16**	**5,90**[a]	**>**
M27[d]	Anz. falsch eher irrel. bew. eher rel. Dok.	0,67	0,47	0,62	0,53	0,57	0,66	0,49	
M26[b]	Anz. falsch eher irrel. bew. Dok.	1,53	1,22	1,35	1,41	**1,17**	**< 1,84 >**	**1,12**[a]	
Z01-log	Durchschn. Betrachtungsz. aller Dok.	**3,64**	**3,46**[a]	3,50	3,61	**3,75**	**> 3,51**	**3,39**[a]	**<**

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Fortsetzung auf nächster Seite

[b] Warnung bei Posthoc-Test für Positionseffekt.

[c] Mittelwertuntschied für Positionseffekt nicht testbar.

[d] Für M27 ist in der Tendenz auch eine Wechselwirkung zwischen Systemleistung und Erwartungshaltung nachweisbar.

Tab. E.26 (Fortsetzung)

ID	Beschreibung	System		Erwartung		Position			
		S_G	S_S	E_H	E_N	P_1	P_2	P_3	
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	3,78	3,70	3,66	3,82	**3,93**	> **3,68**		**3,60**[a] <
Z08	Durchschn. Betrachtungsz. rel. Dok.	57,36	48,89	48,89	57,36	**61,65**	> **50,79**		**46,93**[a] <
Z08-log[b]	Durchschn. Betrachtungsz. rel. Dok.	3,71	3,60	3,59	3,72	**3,85**	> **3,63**	> **3,48**[a]	<
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	**57,98**	**44,98**[a]	47,85	55,11	**60,09**	> **48,80**		**45,55**[a] <
		57,98	**45,87**[a]	47,40	56,45	**62,49**	> **48,16**		**45,12**[a] <
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	**62,33**	**52,00**[a]	50,11	64,22	**67,59**	> **53,73**		**50,17**[a] <
Z11[b]	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	**62,13**	**51,90**[a]	52,00	62,03	**68,87**	> **51,44**		**50,73**[a] <
Z11-log	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	3,79	3,69	3,67	3,81	**3,93**	> **3,69**		**3,60**[a] <
Z23	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	57,18	53,89	49,07	62,00	**64,54**	> **53,23**		**48,83**[a] <
V05[b]	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	**0,30**	**0,18**[a]	0,22	0,26	**0,23**		**0,29** > **0,19**[a]	
		0,30	**0,19**[a]	0,21	0,27	**0,24**		**0,29** > **0,19**[a]	
S05-log	Zeit zum ersten richtig rel. bew. Dok.	4,90	4,77	4,78	4,88	**4,99**		**4,91** > **4,60**[a]	<

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

[b] Warnung bei Posthoc-Test für Positionseffekt.

[c] Mittelwertuntschied für Positionseffekt nicht testbar.

[d] Für M27 ist in der Tendenz auch eine Wechselwirkung zwischen Systemleistung und Erwartungshaltung nachweisbar.

Tab. E.27.: Weitere signifikante Ergebnisse der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung ohne Positionseffekt in Stichprobe SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier Stichproben nachweisbar sind.

ID	Beschreibung	System		Erwartung		Position		
		S_G	S_S	E_H	E_N	P_1	P_2	P_3
M05	Anz. aufg. irrel. Dok.	**1,47**[a]	**2,76**	2,13	2,11	1,90	2,12	2,34
M14	Anz. richtig bew. Dok.	**3,16**	**5,60**[a]	4,05	4,71	3,78	4,42	4,94
M15	Anz. richtig irrel. bew. Dok.	**1,12**	**1,89**[a]	1,47	1,54	1,39	1,46	1,67
M22	Anz. aufg. irrel. Dok. (4-st.)	**0,58**[a]	**1,04**	0,70	0,92	0,79	0,67	0,97
M27	Anz. falsch eher irrel. bew. eher rel. Dok.	**0,66**	**0,46**[a]	0,57	0,54	0,56	0,67	0,44
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	**5,30**	**5,82**[a]	5,69	5,43	5,56	5,27	5,86
V01	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	**0,15**[a]	**0,28**	0,20	0,23	0,20	0,22	0,23
V02	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	**0,85**[a]	**0,72**	0,79	0,77	0,79	0,78	0,77
V06	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	**0,69**	**0,46**[a]	0,59	0,55	0,61	0,60	0,51
V13	Anz. richtig bew. Dok. / Anz. aufg. Dok.	**0,39**	**0,60**[a]	0,46	0,52	0,46	0,49	0,53
V14	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	**0,11**	**0,21**[a]	0,15	0,17	0,15	0,16	0,16
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	**0,30**	**0,53**[a]	0,39	0,44	0,39	0,37	0,48
V36	Anz. aufg. irrel. Dok. / Anz. aufg. Dok. (4-st.)	**0,044**[a]	**0,097**	0,067	0,073	0,076	0,064	0,07
V44	Anz. falsch eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	0,62	0,66	**0,69**	**0,59**[a]	0,61	0,62	0,68
V54	Anz. irrel. bew. Dok. / Anz. aufg. Dok. (4-st.)	0,20	0,21	**0,17**[a]	**0,24**	0,22	0,20	0,20

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Fortsetzung auf nächster Seite

Tab. E.27 (Fortsetzung)

		System		Erwartung		Position		
ID	**Beschreibung**	**S_G**	**S_S**	**E_H**	**E_N**	**P_1**	**P_2**	**P_3**
V58	Anz. richtig eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	0,37	0,35	**0,30**	**0,42**[a]	0,42	0,35	0,32

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Tab. E.28.: Weitere signifikante Ergebnisse der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung ohne Positionseffekt in Stichprobe SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System		Erwartung		Position		
ID	**Beschreibung**	**S_G**	**S_S**	**E_H**	**E_N**	**P_1**	**P_2**	**P_3**
M05	Anz. aufg. irrel. Dok.	**1,48**[a]	**3,10**	2,30	2,28	1,97	2,29	2,61
M20	Anz. aufg. eher irrel. Dok.	**0,92**[a]	**1,82**	1,47	1,27	1,00	1,54	1,57
M48	Anz. richtig rel. bew. Dok. (letzte Suche) (4-st.)	1,30	1,57	**1,74**[a]	**1,12**	1,29	1,40	1,61
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	5,29	5,70	**5,84**[a]	**5,15**	5,50	5,19	5,79
Z24	Durchschn. Betrachtungsz. richtig bew. Dok. (4-st.)	**55,36**	**37,23**[a]	44,02	48,57	53,49	45,03	40,36
V01	Anz. aufg. irrel. Dok. / Anz. aufg. Dok.	**0,15**[a]	**0,30**	0,23	0,23	0,21	0,23	0,24
V02	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	**0,85**[a]	**0,71**	0,78	0,77	0,79	0,78	0,76
V06	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	**0,69**	**0,45**[a]	0,58	0,57	0,60	0,61	0,51
V14	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	**0,11**	**0,22**[a]	0,15	0,18	0,16	0,17	0,17
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	**0,31**	**0,56**[a]	0,40	0,46	0,41	0,39	0,49
V29	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	**0,65**	**0,75**[a]	0,73	0,68	0,70	0,65	0,76

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Tab. E.29.: Weitere signifikante Ergebnisse der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerzufriedenheit ohne Positionseffekt in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier Stichproben nachweisbar sind.

		System		Erwartung		Position		
ID	**Beschreibung**	**S_G**	**S_S**	**E_H**	**E_N**	**P_1**	**P_2**	**P_3**
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	3,28	3,17	**3,47**[a]	**2,97**	3,37	3,11	3,18
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	3,29	3,30	**3,52**[a]	**3,07**	3,37	3,22	3,30
F04	Liefert die Suchmaschine genügend Information?	3,38	3,03	**3,56**[a]	**2,85**	3,36	3,21	3,04
F05	Ist die Suchmaschine präzise?	3,05	2,73	**3,36**[a]	**2,42**	2,90	2,83	2,93

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

Fortsetzung auf nächster Seite

[b] Warnung bei Posthoc-Test für Positionseffekt.

[c] Entspricht auch den Skalen SK15-M und SK19-M.

[d] Entspricht auch der Skala SK13-M.

Tab. E.29 (Fortsetzung)

		System		Erwartung		Position		
ID	**Beschreibung**	S_G	S_S	E_H	E_N	P_1	P_2	P_3
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	3,08	2,85	**3,34**[a]	**2,59**	3,04	2,89	2,97
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	3,26	3,05	**3,51**[a]	**2,80**	3,29	3,04	3,12
F08	Ist die Suchmaschine benutzerfreundlich?	3,99	3,72	**4,23**[a]	**3,48**	3,89	3,82	3,86
F09[b]	Ist die Suchmaschine einfach zu bedienen?	4,38	4,50	4,53	4,35	**4,69**[a]	**> 4,33**	**4,29 <**
F12	Ist die Suchmaschine erfolgreich?	3,36	3,39	**3,69**[a]	**3,06**	3,59	3,25	3,28
F13	Sind Sie mit der Suchmaschine zufrieden?	3,43	3,12	**3,56**[a]	**2,98**	3,60	3,11	3,11
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	3,34	3,52	**3,71**[a]	**3,15**	3,74	3,17	3,39
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	3,18	2,78	**3,32**[a]	**2,63**	3,18	2,87	2,87
F22	Ich bin mit den Suchergebnissen zufrieden.	3,31	3,26	**3,64**[a]	**2,93**	3,46	3,29	3,10
F23	Ich bin mit meiner Suchleistung zufrieden.	3,31	3,46	**3,63**[a]	**3,14**	3,53	3,25	3,37
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	3,05	3,03	**3,38**[a]	**2,69**	3,10	2,90	3,11
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	2,83	2,69	**3,21**[a]	**2,31**	2,90	2,68	2,71
SK01-M	Genauigkeit	2,99	3,03	**3,34**[a]	**2,68**	3,10	2,96	2,97
SK02-M	Inhalt	3,31	3,11	**3,55**[a]	**2,87**	3,39	3,07	3,17
SK03-M	Benutzerfreundlichkeit	3,65	3,35	**3,83**[a]	**3,17**	3,65	3,33	3,52
SK04-M	Suche	3,23	3,18	**3,49**[a]	**2,92**	3,39	3,12	3,09
SK08-M	Suche	3,22	3,13	**3,47**[a]	**2,89**	3,31	3,10	3,11
SK09-M	Benutzerfreundlichkeit	**3,55**[a]	**3,20**	**3,62**[a]	**3,13**	3,47	3,29	3,37
		3,50[a]	**3,20**	**3,68**[a]	**3,02**	3,44	3,21	3,40
SK11-M[c]	Eigenleistung	3,08	3,01	**3,26**[a]	**2,83**	3,10	2,98	3,06
SK12-M	Suchergebnis	3,33	3,09	**3,49**[a]	**2,93**	3,34	3,11	3,18
SK14-M	Suche	3,29	3,09	**3,48**[a]	**2,90**	3,36	3,02	3,20
SK16-M	Aufgabe	3,61	3,52	**3,81**[a]	**3,32**	3,86	3,37	3,47
SK17-M	Suche	3,17	3,04	**3,45**[a]	**2,77**	3,17	3,04	3,12
SK-A	Accuracy (EUCS)	3,09	2,83	**3,31**[a]	**2,61**	3,06	2,87	2,95
SK-C	Content (EUCS)	3,30	3,19	**3,57**[a]	**2,92**	3,35	3,18	3,21
SK-E[d]	Ease of Use (EUCS)	4,17	4,09	**4,39**[a]	**3,87**	4,28	4,05	4,06
SK-E-88	EUCS-Skala-1988	3,41	3,28	**3,64**[a]	**3,05**	3,53	3,25	3,25
SK-E-13	EUCS-Skala-2013	3,43	3,25	**3,63**[a]	**3,05**	3,45	3,23	3,34
SK-E-09	EUCS-Skala-2009	3,25	3,18	**3,62**[a]	**2,81**	3,26	3,19	3,20
SK-G-13	Gesamtskala-2013	3,24	3,11	**3,47**[a]	**2,88**	3,28	3,08	3,16
SK-K	Kriteriumsskala	3,32	3,24	**3,66**[a]	**2,89**	3,62	3,12	3,10
SK-Z-13	Zusatzskala-2013	3,19	3,27	**3,53**[a]	**2,92**	3,37	3,13	3,18
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	3,33	3,12	**3,65**[a]	**2,81**	3,39	3,24	3,06
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	3,18	3,01	**3,29**[a]	**2,90**	3,16	3,03	3,10
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	3,04	2,87	**3,30**[a]	**2,61**	3,03	2,89	2,94
E06-M	Erwartungsskala	3,41	3,09	**3,47**[a]	**3,02**	3,37	3,19	3,18

[a] Dieser Mittelwert entspricht der besseren Benutzerleistung.

[b] Warnung bei Posthoc-Test für Positionseffekt.

[c] Entspricht auch den Skalen SK15-M und SK19-M.

[d] Entspricht auch der Skala SK13-M.

Tab. E.30.: Weitere signifikante Ergebnisse der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerzufriedenheit ohne Positionseffekt in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

		System		Erwartung		Position		
ID	**Beschreibung**	S_G	S_S	E_H	E_N	P_1	P_2	P_3
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	3,32	3,22	**3,64**[a]	**2,91**	3,36	3,11	3,35
F12	Ist die Suchmaschine erfolgreich?	3,41	3,34	**3,77**[a]	**2,98**	3,59	3,22	3,31
SK11-M[b]	Eigenleistung	2,99	3,07	**3,39**[a]	**2,67**	3,11	2,90	3,08
SK-E[c]	Ease of Use (EUCS)	4,13	4,07	**4,44**[a]	**3,76**	4,23	3,99	4,07
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	3,21	3,01	**3,36**[a]	**2,86**	3,20	3,00	3,12

[a] Dieser Mittelwert entspricht der höheren Benutzerzufriedenheit.
[b] Entspricht auch den Skalen SK15-M und SK19-M.
[c] Entspricht auch der Skala SK13-M.

E.4.3. Mittelwerte der Interaktionen

In diesem Abschnitt sind die Mittelwerte der im Rahmen der dynamischen Analysen signifikanten Interaktionen zusammengefasst. Im Gegensatz zur Benutzerleistung lassen sich im Kontext der Zufriedenheit keine tendenziell oder eindeutig signifikanten Interaktionen nachweisen. Um dennoch einen Eindruck von der dynamischen Abhängigkeit der Zufriedenheit zu gewinnen, werden im Rahmen der Hauptauswertung auch Zufriedenheitsvariablen in die Betrachtung mit einbezogen, für die sich nur in einzelnen Fällen signifikante Wechselwirkungseffekte zeigen. Dies muss bei der Betrachtung der in diesem Abschnitt aufgeführten Tabellen berücksichtigt werden.

Tab. E.31.: Mittelwerte signifikanter Interaktionen zwischen System und Erwartung im Rahmen der klassischen Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerleistung in SP_A.

		Interaktion: System-Erwartung			
ID	**Beschreibung**	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
Z07	Durchschn. Betrachtungsz. rel. bew. Dok.	**47,60**	**73,34**	**73,82**	**48,62**

Tab. E.32.: Mittelwerte signifikanter Interaktionen zwischen System und Aufgabenposition im Rahmen der klassischen Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerleistung in SP_A

		Interaktion: System-Position					
ID	**Beschreibung**	$I_{G,P1}$	$I_{G,P2}$	$I_{G,P3}$	$I_{S,P1}$	$I_{S,P2}$	$I_{S,P3}$
B05	Durchschn. Bew. rel. Dok. (erste Suche)	**5,08**	**5,08**	**5,60**	**5,58**	**5,58**	**4,98**
S04	Suchdauer	**524,50**	**488,36**	**455,40**	**489,74**	**535,97**	**434,40**

Tab. E.33.: Mittelwerte signifikanter Interaktionen zwischen System und Aufgabenposition im Rahmen der klassischen Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerleistung in SP_B

		Interaktion: System-Position					
ID	**Beschreibung**	$I_{G,P1}$	$I_{G,P2}$	$I_{G,P3}$	$I_{S,P1}$	$I_{S,P2}$	$I_{S,P3}$
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,41**	**0,50**	**0,32**	**0,38**	**0,44**	**0,42**
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,59**	**0,50**	**0,68**	**0,63**	**0,57**	**0,59**
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	**0,55**	**0,49**	**0,66**	**0,57**	**0,54**	**0,52**

Tab. E.34.: Mittelwerte signifikanter Interaktionen zwischen System, Erwartung und Aufgabenposition im Rahmen der Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerleistung in SP_A. Die Mittelwerte der Dreifachwechselwirkung sind aufgeteilt in gut und schlechte Systemleistung.

			Erwartung-Position					
ID	**Beschreibung**	**System**	$I_{H,P1}$	$I_{H,P2}$	$I_{H,P3}$	$I_{N,P1}$	$I_{N,P2}$	$I_{N,P3}$
M27	Anz. falsch eher irrel. bew. eher rel. Dok.	S_G	**0,66**	**1,00**	**0,52**	**0,59**	**0,48**	**0,66**
		S_S	**0,59**	**0,48**	**0,28**	**0,45**	**0,72**	**0,28**

Tab. E.35.: Mittelwerte signifikanter Interaktionen zwischen System, Erwartung und Aufgabenposition im Rahmen der Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerleistung in SP_B. Die Mittelwerte der Dreifachwechselwirkung sind aufgeteilt in gut und schlechte Systemleistung.

			Erwartung-Position					
ID	**Beschreibung**	**System**	$I_{H,P1}$	$I_{H,P2}$	$I_{H,P3}$	$I_{N,P1}$	$I_{N,P2}$	$I_{N,P3}$
M27	Anz. falsch eher irrel. bew. eher rel. Dok.	S_G	**0,65**	**1,15**	**0,70**	**0,55**	**0,35**	**0,65**
		S_S	**0,60**	**0,40**	**0,25**	**0,40**	**0,85**	**0,40**

Tab. E.36.: Mittelwerte signifikanter Interaktionen zwischen System und Erwartung im Rahmen der klassischen Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerzufriedenheit in SP_A.

		Interaktion: System-Erwartung			
ID	**Beschreibung**	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	**3,31**	**3,15**	**3,24**	**2,39**
SK03-M	Benutzerfreundlichkeit	**3,76**	**3,47**	**3,92**	**2,81**
SK09-M	Benutzerfreundlichkeit	**3,43**	**3,48**	**3,67**	**2,83**
SK12-M	Suchergebnis	**3,41**	**3,24**	**3,56**	**2,62**
SK-G-13	Gesamtskala-2013	**3,37**	**3,27**	**3,39**	**2,72**

Tab. E.37.: Mittelwerte signifikanter Interaktionen zwischen System und Aufgabenposition im Rahmen der klassischen Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerzufriedenheit in SP_A

		Interaktion: System-Position					
ID	**Beschreibung**	$I_{G,P1}$	$I_{G,P2}$	$I_{G,P3}$	$I_{S,P1}$	$I_{S,P2}$	$I_{S,P3}$
F04	Liefert die Suchmaschine genügend Information?	**3,44**	**3,44**	**3,50**	**3,47**	**2,86**	**2,78**
F08	Ist die Suchmaschine benutzerfreundlich?	**3,75**	**3,83**	**4,00**	**4,00**	**3,58**	**3,61**
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	**2,36**	**2,61**	**2,97**	**3,11**	**2,64**	**2,83**
F23	Ich bin mit meiner Suchleistung zufrieden.	**3,19**	**2,86**	**3,42**	**3,44**	**3,36**	**3,08**
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	**2,78**	**2,86**	**2,86**	**3,03**	**2,50**	**2,56**

Tab. E.38.: Mittelwerte signifikanter Interaktionen zwischen Erwartung und Aufgabenposition im Rahmen der klassischen Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerzufriedenheit in SP_A.

		Interaktion: Erwartung-Position					
ID	**Beschreibung**	$I_{H,P1}$	$I_{H,P2}$	$I_{H,P3}$	$I_{N,P1}$	$I_{N,P2}$	$I_{N,P3}$
F04	Liefert die Suchmaschine genügend Information?	**3,64**	**3,31**	**3,67**	**3,28**	**3,00**	**2,61**
F08	Ist die Suchmaschine benutzerfreundlich?	**4,22**	**4,08**	**4,47**	**3,83**	**3,81**	**3,50**
F13	Sind Sie mit der Suchmaschine zufrieden?	**3,94**	**3,11**	**3,47**	**3,17**	**3,17**	**2,72**
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	**3,36**	**3,08**	**3,69**	**2,83**	**2,72**	**2,53**
SK07-M	Benutzerfreundlichkeit	**3,65**	**3,18**	**3,51**	**3,14**	**3,22**	**2,96**
SK09-M	Benutzerfreundlichkeit	**3,58**	**3,50**	**3,79**	**3,36**	**3,07**	**2,95**
SK-E	Ease of Use (EUCS)	**4,51**	**4,26**	**4,41**	**4,06**	**3,84**	**3,70**
SK-T	Timeliness (EUCS)	**3,90**	**3,62**	**3,94**	**3,65**	**3,39**	**2,97**
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	**3,78**	**3,47**	**3,58**	**3,00**	**3,03**	**2,36**
E06-M	Erwartungsskala	**3,51**	**3,24**	**3,51**	**3,21**	**3,08**	**2,81**

Tab. E.39.: Mittelwerte signifikanter Interaktionen zwischen Erwartung und Aufgabenposition im Rahmen der Varianzanalyse zur Untersuchung der dynamischen Entwicklung der Benutzerzufriedenheit in SP_B.

		Interaktion: Erwartung-Position					
ID	**Beschreibung**	$I_{H,P1}$	$I_{H,P2}$	$I_{H,P3}$	$I_{N,P1}$	$I_{N,P2}$	$I_{N,P3}$
SK-E	Ease of Use (EUCS)	**4,46**	**4,30**	**4,50**	**4,04**	**3,66**	**3,62**

E.4.4. Teststatistiken der dynamischen Analyse

Die in diesem Abschnitt dargestellten Tabellen enthalten die Teststatistiken zu den im Rahmen des dritten Experiments durchgeführten dynamischen Analysen. Die Tabellen zeigen neben der jeweiligen Stichprobengröße, der zugrundeliegenden Analysevariante und den Freiheitsgraden, den jeweils erreichten F-Wert sowie das zugehörige Signifikanzniveau.

Tab. E.40.: Teststatistik der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**Art**	**F**	**p**
M01	Anz. aufg. Dok.	nein	116	R	-	0,62	0,43	0,05	0,82	12,58	**$4{\cdot}10^{-6}$**	I_{SE}	0,31	0,58
												I_{SP}	1,33	0,27
												I_{EP}	0,20	0,82
												I_{SEP}	0,27	0,77
M05	Anz. aufg. irrel. Dok.	nein	116	R	-	33,23	**$1{,}1{\cdot}10^{-8}$**	0,45	0,50	2,73	0,066	I_{SE}	0,0056	0,94
												I_{SP}	1,15	0,32
												I_{EP}	0,31	0,73
												I_{SEP}	0,30	0,74
M06	Anz. aufg. rel. Dok.	ja	72	R	-	1,70	0,19	2,74	0,098	7,28	**0,00073**	I_{SE}	0,023	0,88
												I_{SP}	0,12	0,88
												I_{EP}	0,12	0,88
												I_{SEP}	0,24	0,78
M07	Anz. falsch irrel. bew. Dok.	nein	116	R	-	7,45	**0,0065**	0,031	0,86	5,36	**0,0048**	I_{SE}	1,72	0,19
												I_{SP}	0,29	0,75
												I_{EP}	0,98	0,37
												I_{SEP}	0,14	0,87
M10	Anz. rel. bew. Dok.	nein	116	R	-	0,81	0,37	1,36	0,24	6,07	**0,0024**	I_{SE}	0,13	0,72
												I_{SP}	0,27	0,77
												I_{EP}	0,60	0,55
												I_{SEP}	0,27	0,76
M14	Anz. richtig bew. Dok.	ja	72	R	-	12,74	**0,00038**	1,14	0,29	1,42	0,24	I_{SE}	3,27	0,071
												I_{SP}	0,61	0,54
												I_{EP}	0,30	0,74
												I_{SEP}	0,41	0,66
M15	Anz. richtig irrel. bew. Dok.	ja	72	R	-	15,82	**$7{,}5{\cdot}10^{-5}$**	0,76	0,38	0,76	0,47	I_{SE}	0,25	0,62
												I_{SP}	0,64	0,53
												I_{EP}	0,30	0,74
												I_{SEP}	0,18	0,84

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.40 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
M20	Anz. aufg. eher irrel. Dok.	nein	116	R	-	32,27	$\mathbf{1{,}8{\cdot}10^{-8}}$	1,85	0,17	8,37	**0,00025**	I_{SE}	0,061	0,80
												I_{SP}	1,88	0,15
												I_{EP}	0,071	0,93
												I_{SEP}	0,23	0,79
M22	Anz. aufg. irrel. Dok. (4-st.)	ja	72	R	-	10,48	**0,0012**[a]	1,04	0,31	1,55	0,21	I_{SE}	0,029	0,86
												I_{SP}	3,29	0,038
												I_{EP}	0,18	0,83
												I_{SEP}	0,055	0,95
M26	Anz. falsch eher irrel. bew. Dok.	nein	116	R	-	4,04	0,045	0,56	0,46	7,63	**0,00051**	I_{SE}	2,77	0,096
												I_{SP}	0,71	0,49
												I_{EP}	0,56	0,57
												I_{SEP}	0,63	0,53
M27	Anz. falsch eher irrel. bew. eher rel. Dok.	nein	116	R	-	5,85	**0,016**[a]	0,12	0,73	1,85	0,16	I_{SE}	0,75	0,39
												I_{SP}	1,46	0,23
												I_{EP}	1,14	0,32
												I_{SEP}	2,57	0,077
M37	Anz. rel. bew. Dok. (4-st.)	nein	116	R	-	1,11	0,29	0,31	0,58	8,39	**0,00024**	I_{SE}	0,31	0,58
												I_{SP}	1,12	0,33
												I_{EP}	1,30	0,27
												I_{SEP}	2,71	0,067
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	nein	96	R	-	7,12	**0,0078**[a]	2,00	0,16	3,53	0,03	I_{SE}	0,81	0,37
												I_{SP}	0,13	0,87
												I_{EP}	0,71	0,49
												I_{SEP}	0,20	0,82
Z01	Durchschn. Betrachtungsz. aller Dok.	ja	72	R	-	0,68	0,41	0,047	0,83	5,15	**0,006**	I_{SE}	0,46	0,50
												I_{SP}	0,47	0,63
												I_{EP}	0,26	0,77
												I_{SEP}	0,19	0,82

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.40 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
Z01-log	Durchschn. Betrachtungsz. aller Dok.	nein	116	R	-	1,53	0,22	0,76	0,38	9,82	$\mathbf{6{\cdot}10^{-5}}$	I_{SE}	0,41	0,52
												I_{SP}	0,67	0,51
												I_{EP}	0,41	0,67
												I_{SEP}	1,34	0,26
Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	nein	88	R	-	4,76	0,029	0,099	0,75	5,22	**0,0055**	I_{SE}	0,037	0,85
												I_{SP}	0,64	0,53
												I_{EP}	0,0078	0,99
												I_{SEP}	1,68	0,19
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	nein	108	R	-	0,52	0,47	$1{,}5{\cdot}10^{-5}$	1,00	9,93	$\mathbf{5{,}4{\cdot}10^{-5}}$	I_{SE}	4,19	0,041
												I_{SP}	0,14	0,87
												I_{EP}	1,40	0,25
												I_{SEP}	0,84	0,43
Z08	Durchschn. Betrachtungsz. rel. Dok.	nein	116	R	-	1,16	0,28	0,24	0,62	11,75	$\mathbf{9{\cdot}10^{-6}}$	I_{SE}	3,27	0,071
												I_{SP}	0,50	0,61
												I_{EP}	1,62	0,20
												I_{SEP}	0,12	0,89
Z08-log	Durchschn. Betrachtungsz. rel. Dok.	nein	116	R	-	0,49	0,48	0,21	0,65	11,40	$\mathbf{1{,}3{\cdot}10^{-5}}$	I_{SE}	0,20	0,66
												I_{SP}	0,16	0,86
												I_{EP}	1,35	0,26
												I_{SEP}	1,02	0,36
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	nein	112	R	-	0,61	0,44	0,0071	0,93	7,24	**0,00076**	I_{SE}	1,43	0,23
												I_{SP}	0,73	0,48
												I_{EP}	0,50	0,61
												I_{SEP}	0,052	0,95
Z09-log	Durchschn. Betrachtungsz. richtig bew. Dok.	ja	72	K[b]	1/68	5,47	0,022	0,074	0,79	18,40	$\mathbf{8{,}5{\cdot}10^{-8}}$[a]	I_{SE}	1,13	0,29
												I_{SP}	0,92	0,40
												I_{EP}	0,50	0,61
												I_{SEP}	0,23	0,80

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.40 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	nein	108	R	-	1,87	0,17	0,24	0,62	7,12	**0,00085**	I_{SE}	3,92	0,048
												I_{SP}	1,21	0,30
												I_{EP}	1,43	0,24
												I_{SEP}	0,21	0,81
Z11-log	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	nein	108	K[b]	1/104	0,22	0,64	0,0066	0,94	25,14	$\mathbf{5{,}4{\cdot}10^{-10}}$	I_{SE}	5,67	0,019
												I_{SP}	0,92	0,40
												I_{EP}	1,67	0,19
												I_{SEP}	0,42	0,64
Z14-log	Durchschn. Betrachtungsz. eher rel. bew. Dok.	nein	68	R	-	0,78	0,38	0,41	0,52	4,74	**0,009**	I_{SE}	1,93	0,17
												I_{SP}	1,66	0,19
												I_{EP}	0,0073	0,99
												I_{SEP}	1,72	0,18
Z15	Durchschn. Betrachtungsz. eher rel. Dok.	nein	76	R	-	4,03	0,045	0,015	0,90	4,95	**0,0072**	I_{SE}	0,099	0,75
												I_{SP}	0,37	0,69
												I_{EP}	0,088	0,92
												I_{SEP}	1,27	0,28
Z23	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	nein	116	R	-	0,74	0,39	0,0021	0,96	7,24	**0,00075**	I_{SE}	0,56	0,45
												I_{SP}	2,42	0,09
												I_{EP}	1,58	0,21
												I_{SEP}	0,078	0,92
Z23-log	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	ja	72	R	-	2,25	0,13	1,91	0,17	5,91	**0,0028**	I_{SE}	0,021	0,88
												I_{SP}	0,30	0,74
												I_{EP}	0,84	0,43
												I_{SEP}	0,16	0,86
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	116	R	-	62,19	$\mathbf{8{,}1{\cdot}10^{-15}}$	4,02	0,045	0,56	0,57	I_{SE}	0,074	0,79
												I_{SP}	0,57	0,57
												I_{EP}	1,26	0,28
												I_{SEP}	0,54	0,58

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.40 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
V02	Anz. aufg. rel. Dok. / Anz. aufg. Dok.	nein	116	R	-	63,06	**$5{,}3{\cdot}10^{-15}$**	2,27	0,13	0,94	0,39	I_{SE}	0,058	0,81
												I_{SP}	0,73	0,48
												I_{EP}	1,21	0,30
												I_{SEP}	0,78	0,46
V03	Anz. aufg. rel. Dok. / Anz. rel. Dok. im Korpus	ja	72	R	-	0,84	0,36	1,48	0,22	7,45	**0,00062**	I_{SE}	0,52	0,47
												I_{SP}	0,30	0,74
												I_{EP}	0,63	0,53
												I_{SEP}	1,95	0,14
V04	Anz. aufg. rel. Dok. / Anz. zurückgeg. rel. Dok.	ja	72	R	-	1,04	0,31	0,37	0,54	7,20	**0,00078**	I_{SE}	0,64	0,42
												I_{SP}	0,13	0,88
												I_{EP}	0,76	0,47
												I_{SEP}	0,83	0,44
V05	Anz. falsch irrel. bew. Dok. / Anz. aufg. Dok.	nein	116	R	-	13,25	**0,00029**	0,93	0,33	4,48	**0,012**	I_{SE}	0,028	0,87
												I_{SP}	1,18	0,31
												I_{EP}	1,09	0,34
												I_{SEP}	0,14	0,87
						12,95	**0,00034**	0,28	0,60	6,59	**0,0014**	I_{SE}	0,12	0,73
												I_{SP}	0,61	0,55
												I_{EP}	0,80	0,45
												I_{SEP}	0,047	0,95
V06	Anz. falsch irrel. bew. Dok. / Anz. irrel. bew. Dok.	nein	88	R	-	29,77	**$6{,}1{\cdot}10^{-8}$**	0,98	0,32	1,46	0,23	I_{SE}	0,87	0,35
												I_{SP}	1,00	0,37
												I_{EP}	0,70	0,50
												I_{SEP}	1,18	0,31
V13	Anz. richtig bew. Dok. / Anz. aufg. Dok.	ja	72	R	-	16,32	**$5{,}8{\cdot}10^{-5}$**	0,90	0,34	0,34	0,71	I_{SE}	0,36	0,55
												I_{SP}	1,08	0,34
												I_{EP}	0,21	0,81
												I_{SEP}	0,0068	0,99

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.40 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
V14	Anz. richtig irrel. bew. Dok. / Anz. aufg. Dok.	nein	116	R	-	32,16	$\mathbf{1{,}9{\cdot}10^{-8}}$	2,54	0,11	0,43	0,65	I_{SE}	0,077	0,78
												I_{SP}	2,00	0,14
												I_{EP}	0,19	0,82
												I_{SEP}	0,12	0,89
V17	Anz. richtig irrel. bew. Dok. / Anz. irrel. bew. Dok.	nein	88	R	-	36,18	$\mathbf{2{,}5{\cdot}10^{-9}}$	1,65	0,20	2,44	0,088	I_{SE}	0,40	0,53
												I_{SP}	1,13	0,32
												I_{EP}	1,01	0,36
												I_{SEP}	2,08	0,13
V29	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	nein	108	R	-	3,42	0,065	0,035	0,85	5,52	**0,0041**	I_{SE}	0,28	0,60
												I_{SP}	1,24	0,29
												I_{EP}	0,81	0,45
												I_{SEP}	1,13	0,33
V32/BR	Anz. richtig rel. bew. Dok. / Anz. rel. Dok. im Korpus	nein	108	R	-	0,31	0,58	2,09	0,15	5,72	**0,0034**	I_{SE}	0,14	0,71
												I_{SP}	0,048	0,95
												I_{EP}	0,50	0,61
												I_{SEP}	0,27	0,76
V33	Anz. richtig rel. bew. Dok. / Anz. zurückgeg. rel. Dok.	nein	108	R	-	0,69	0,41	0,33	0,57	5,29	**0,0052**	I_{SE}	1,37	0,24
												I_{SP}	0,015	0,98
												I_{EP}	0,81	0,45
												I_{SEP}	0,23	0,79
V36	Anz. aufg. irrel. Dok. / Anz. aufg. Dok. (4-st.)	ja	72	R	-	14,12	**0,00018**	0,024	0,88	0,22	0,81	I_{SE}	0,0095	0,92
												I_{SP}	0,28	0,75
												I_{EP}	0,25	0,78
												I_{SEP}	0,24	0,79
V38	Anz. aufg. rel. Dok. / Anz. rel. Dok. im Korpus (4-st.)	ja	72	R	-	0,012	0,91	2,84	0,092	4,70	**0,0093**[a]	I_{SE}	0,46	0,50
												I_{SP}	0,14	0,87
												I_{EP}	0,45	0,64
												I_{SEP}	0,0083	0,99

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.40 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
V44	Anz. falsch eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	nein	68	R	-	2,16	0,14	6,95	**0,0085**	0,33	0,72	I_{SE}	0,057	0,81
												I_{SP}	0,52	0,60
												I_{EP}	1,23	0,29
												I_{SEP}	0,86	0,42
V54	Anz. irrel. bew. Dok. / Anz. aufg. Dok. (4-st.)	ja	72	R	-	0,19	0,67	6,50	**0,011**[a]	0,36	0,70	I_{SE}	0,86	0,35
												I_{SP}	1,51	0,22
												I_{EP}	0,24	0,78
												I_{SEP}	0,42	0,66
V58	Anz. richtig eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	nein	68	R	-	1,40	0,24	10,56	**0,0012**	0,70	0,49	I_{SE}	0,41	0,52
												I_{SP}	0,071	0,93
												I_{EP}	0,86	0,43
												I_{SEP}	1,65	0,19
S04	Suchdauer	nein	116	R	-	0,20	0,66	0,24	0,62	6,59	**0,0014**	I_{SE}	0,20	0,66
												I_{SP}	4,15	0,016
												I_{EP}	0,72	0,49
												I_{SEP}	1,65	0,19
S05-log	Zeit zum ersten richtig rel. bew. Dok.	nein	104	K[b]	1/100	0,15	0,70	0,00012	0,99	19,24	$\mathbf{2{,}3{\cdot}10^{-8}}$	I_{SE}	2,41	0,12
												I_{SP}	2,67	0,072
												I_{EP}	0,50	0,61
												I_{SEP}	0,18	0,83

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

Tab. E.41.: Teststatistik der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**Art**	**F**	**p**
M01	Anz. aufg. Dok.	nein	80	R	-	3,65	0,056	3,37	0,067	7,75	**0,00046**	I_{SE}	0,053	0,82
												I_{SP}	0,77	0,47
												I_{EP}	0,28	0,75
												I_{SEP}	0,24	0,79
M05	Anz. aufg. irrel. Dok.	nein	80	R	-	25,82	$\mathbf{4{,}5{\cdot}10^{-7}}$	0,26	0,61	2,21	0,11	I_{SE}	0,054	0,82
												I_{SP}	0,67	0,51
												I_{EP}	0,16	0,85
												I_{SEP}	0,093	0,91
M07	Anz. falsch irrel. bew. Dok.	nein	80	R	-	5,69	**0,017**	0,36	0,55	3,61	**0,028**	I_{SE}	1,04	0,31
												I_{SP}	0,38	0,69
												I_{EP}	0,52	0,60
												I_{SEP}	0,27	0,77
M07	Anz. falsch irrel. bew. Dok.	nein	80	R	-	4,54	**0,033**	0,23	0,63	4,47	**0,012**	I_{SE}	0,83	0,36
												I_{SP}	0,27	0,77
												I_{EP}	0,75	0,47
												I_{SEP}	0,31	0,73
M10	Anz. rel. bew. Dok.	nein	80	R	-	1,59	0,21	2,28	0,13	3,82	**0,022**[a]	I_{SE}	0,03	0,86
												I_{SP}	0,46	0,63
												I_{EP}	0,43	0,65
												I_{SEP}	1,2	0,30
M20	Anz. aufg. eher irrel. Dok.	nein	80	R	-	19,59	$\mathbf{1{,}1{\cdot}10^{-5}}$	0,17	0,68	2,34	0,097	I_{SE}	0,69	0,41
												I_{SP}	0,54	0,59
												I_{EP}	0,15	0,86
												I_{SEP}	0,19	0,83

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.41 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
M26	Anz. falsch eher irrel. bew. Dok.	nein	80	R	-	2,04	0,15	0,15	0,70	6,27	**0,002**	I_{SE}	3,30	0,07
												I_{SP}	0,16	0,85
												I_{EP}	0,34	0,71
												I_{SEP}	0,99	0,37
M27	Anz. falsch eher irrel. bew. eher rel. Dok.	nein	80	R	-	5,41	0,02	0,60	0,44	0,71	0,49	I_{SE}	5,41	**0,02**[a]
												I_{SP}	0,59	0,55
												I_{EP}	0,37	0,69
												I_{SEP}	2,67	0,07
M48	Anz. richtig rel. bew. Dok. (letzte Suche) (4-st.)	nein	80	R	-	2,00	0,16	6,11	**0,014**[a]	1,08	0,34	I_{SE}	0,055	0,81
												I_{SP}	0,0076	0,99
												I_{EP}	0,056	0,95
												I_{SEP}	0,66	0,52
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	nein	68	R[c]	-	3,51	0,061	9,62	**0,002**[a]	1,89	0,15	I_{SE}	0,92	0,34
												I_{SP}	0,43	0,65
												I_{EP}	0,099	0,91
												I_{SEP}	0,38	0,68
Z01-log	Durchschn. Betrachtungsz. aller Dok.	nein	80	R	-	9,20	**0,0025**	1,19	0,28	8,07	**0,00033**	I_{SE}	0,0066	0,94
												I_{SP}	0,19	0,83
												I_{EP}	0,67	0,51
												I_{SEP}	0,68	0,51
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	nein	72	K[b]	1/68	0,34	0,56	1,15	0,29	15,72	$\mathbf{2{,}1{\cdot}10^{-6}}$	I_{SE}	0,87	0,36
												I_{SP}	0,5	0,59
												I_{EP}	1,15	0,32
												I_{SEP}	0,64	0,51
Z08	Durchschn. Betrachtungsz. rel. Dok.	nein	80	R	-	6,88	0,0088	2,64	0,10	8,56	**0,00021**	I_{SE}	0,49	0,48
												I_{SP}	0,32	0,73
												I_{EP}	2,23	0,11
												I_{SEP}	0,067	0,93

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.41 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
Z08-log	Durchschn. Betrachtungsz. rel. Dok.	nein	80	R	-	5,21	0,023	1,46	0,23	10,09	**$4{,}6{\cdot}10^{-5}$**	I_{SE}	0,41	0,52
												I_{SP}	0,045	0,96
												I_{EP}	1,83	0,16
												I_{SEP}	1,17	0,31
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	nein	76	R	-	9,05	**0,0027**	1,48	0,22	4,43	**0,012**	I_{SE}	0,27	0,60
												I_{SP}	1,83	0,16
												I_{EP}	0,89	0,41
												I_{SEP}	0,18	0,83
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	nein	76	R	-	7,60	**0,0059**	2,16	0,14	6,48	**0,0016**	I_{SE}	0,088	0,77
												I_{SP}	0,75	0,47
												I_{EP}	0,75	0,47
												I_{SEP}	0,53	0,59
Z11-log	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	nein	72	K[b]	1/68	0,50	0,48	0,95	0,33	13,70	**$3{,}8{\cdot}10^{-6}$**	I_{SE}	1,25	0,27
												I_{SP}	0,42	0,66
												I_{EP}	2,17	0,12
												I_{SEP}	0,94	0,39
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	nein	72	R	-	8,14	**0,0044**[a]	3,22	0,073	4,37	**0,013**	I_{SE}	2,19	0,14
												I_{SP}	0,73	0,48
												I_{EP}	1,75	0,17
												I_{SEP}	1,06	0,35
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	nein	72	R	-	5,33	**0,021**[a]	2,59	0,11	4,68	**0,0095**	I_{SE}	1,95	0,16
												I_{SP}	0,8	0,45
												I_{EP}	1,42	0,24
												I_{SEP}	0,94	0,39
Z23	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	nein	76	R	-	2,76	0,097	3,44	0,064	5,26	**0,0054**	I_{SE}	0,43	0,51
												I_{SP}	0,58	0,56
												I_{EP}	2,34	0,097
												I_{SEP}	0,34	0,71

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.41 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
Z24	Durchschn. Betrachtungsz. richtig bew. Dok. (4-st.)	nein	56	R	-	9,05	**0,0027**[a]	0,57	0,45	1,77	0,17	I_{SE}	0,044	0,83
												I_{SP}	0,91	0,40
												I_{EP}	0,19	0,83
												I_{SEP}	0,13	0,87
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	80	R	-	68,42	$\mathbf{4{,}4{\cdot}10^{-16}}$	0,077	0,78	0,42	0,66	I_{SE}	0,81	0,37
												I_{SP}	0,43	0,65
												I_{EP}	1,55	0,21
												I_{SEP}	0,19	0,82
V02	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	80	R	-	51,79	$\mathbf{1{,}2{\cdot}10^{-12}}$	1,48	0,22	0,29	0,75	I_{SE}	0,031	0,86
												I_{SP}	0,31	0,73
												I_{EP}	0,91	0,40
												I_{SEP}	0,32	0,73
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	80	R	-	11,64	**0,00067**	1,53	0,22	3,61	**0,027**	I_{SE}	0,15	0,70
												I_{SP}	1,54	0,22
												I_{EP}	0,83	0,44
												I_{SEP}	0,76	0,47
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	80	R	-	8,75	**0,0032**	3,96	0,047	3,74	**0,024**	I_{SE}	1,32	0,25
												I_{SP}	1,39	0,25
												I_{EP}	0,76	0,47
												I_{SEP}	0,56	0,57
V06	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	nein	56	R	-	24,42	$\mathbf{9{,}1{\cdot}10^{-7}}$	0,0019	0,97	2,33	0,098	I_{SE}	2,26	0,13
												I_{SP}	0,54	0,58
												I_{EP}	1,01	0,36
												I_{SEP}	0,3	0,74
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	80	R	-	27,88	$\mathbf{1{,}6{\cdot}10^{-7}}$	2,36	0,12	0,16	0,85	I_{SE}	0,0039	0,95
												I_{SP}	0,99	0,37
												I_{EP}	0,098	0,91
												I_{SEP}	0,37	0,69

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

Fortsetzung auf nächster Seite

Tab. E.41 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
V17	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	nein	56	R	-	23,13	$\mathbf{1{,}7{\cdot}10^{-6}}$	0,94	0,33	1,83	0,16	I_{SE}	0,21	0,65
												I_{SP}	0,33	0,72
												I_{EP}	1,59	0,20
												I_{SEP}	1,58	0,21
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	nein	72	R	-	5,98	**0,015**[a]	1,26	0,26	2,91	0,055	I_{SE}	2,66	0,10
												I_{SP}	1,31	0,27
												I_{EP}	1,09	0,34
												I_{SEP}	2,25	0,11
S05-log	Zeit zum ersten richtig rel. bew. Dok.	nein	68	K	1/64	0,82	0,37	0,53	0,47	11,50	$\mathbf{2{,}6{\cdot}10^{-5}}$	I_{SE}	0,0018	0,97
												I_{SP}	1,15	0,32
												I_{EP}	1,2	0,31
												I_{SEP}	0,12	0,89

[a] Effekt ist in der Tendenz signifikant.
[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.
[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

Tab. E.42.: Teststatistik der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerzufriedenheit in Stichprobe SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**Art**	**F**	**p**
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	ja	72	R	-	0,80	0,37	11,14	**0,00088**	0,64	0,53	I_{SE}	0,06	0,81
												I_{SP}	3,69	0,025
												I_{EP}	0,55	0,58
												I_{SEP}	1,68	0,19

[a] Effekt ist in der Tendenz signifikant.
Fortsetzung auf nächster Seite
[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.
[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.
[d] Entspricht auch den Skalen SK15-M und SK19-M.
[e] Entspricht auch der Skala SK13-M.

Tab. E.42 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	nein	116	R	-	0,024	0,88	11,96	**0,00057**	0,32	0,73	I_{SE}	0,067	0,80
												I_{SP}	0,14	0,87
												I_{EP}	0,33	0,72
												I_{SEP}	0,35	0,71
F04	Liefert die Suchmaschine genügend Information?	ja	72	R	-	7,71	0,0056	20,53	$\mathbf{6{,}6{\cdot}10^{-6}}$[a]	0,96	0,39	I_{SE}	0,19	0,66
												I_{SP}	1,45	0,24
												I_{EP}	0,074	0,93
												I_{SEP}	1,27	0,28
F05	Ist die Suchmaschine präzise?	ja	72	R	-	4,63	0,032	32,67	$\mathbf{1{,}4{\cdot}10^{-8}}$	0,30	0,74	I_{SE}	0,45	0,50
												I_{SP}	0,37	0,69
												I_{EP}	0,20	0,82
												I_{SEP}	0,067	0,94
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	ja	72	R	-	1,32	0,25	16,14	$\mathbf{6{,}3{\cdot}10^{-5}}$	0,042	0,96	I_{SE}	0,021	0,89
												I_{SP}	0,035	0,97
												I_{EP}	0,92	0,40
												I_{SEP}	0,042	0,96
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	ja	72	R	-	1,38	0,24	12,40	**0,00045**	0,57	0,56	I_{SE}	0,86	0,35
												I_{SP}	0,23	0,80
												I_{EP}	0,34	0,71
												I_{SEP}	0,093	0,91
F08	Ist die Suchmaschine benutzerfreundlich?	ja	72	R	-	4,84	0,028	21,43	$\mathbf{4{,}2{\cdot}10^{-6}}$	0,11	0,90	I_{SE}	1,30	0,26
												I_{SP}	0,28	0,76
												I_{EP}	1,49	0,23
												I_{SEP}	0,45	0,64
F09	Ist die Suchmaschine einfach zu bedienen?	ja	72	R	-	0,19	0,66	1,30	0,25	7,20	**0,00079**[a]	I_{SE}	0,38	0,54
												I_{SP}	0,081	0,92
												I_{EP}	0,23	0,79
												I_{SEP}	0,058	0,94

[a] Effekt ist in der Tendenz signifikant.

Fortsetzung auf nächster Seite

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Tab. E.42 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
F12	Ist die Suchmaschine erfolgreich?	nein	116	R	-	0,013	0,91	19,42	**$1{,}2{\cdot}10^{-5}$**	2,95	0,053	I_{SE}	0,82	0,37
												I_{SP}	0,13	0,88
												I_{EP}	0,70	0,50
												I_{SEP}	0,013	0,99
F13	Sind Sie mit der Suchmaschine zufrieden?	ja	72	R	-	5,07	0,025	14,09	**0,00018**	3,78	0,023	I_{SE}	1,18	0,28
												I_{SP}	0,41	0,66
												I_{EP}	0,42	0,66
												I_{SEP}	0,93	0,39
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	ja	72	R	-	0,83	0,36	9,93	**0,0017**	2,03	0,13	I_{SE}	0,0049	0,94
												I_{SP}	0,57	0,56
												I_{EP}	0,41	0,66
												I_{SEP}	0,62	0,54
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	ja	72	R	-	5,15	0,023	14,31	**0,00016**	1,48	0,23	I_{SE}	0,043	0,84
												I_{SP}	0,075	0,93
												I_{EP}	0,81	0,45
												I_{SEP}	0,68	0,51
F22	Ich bin mit den Suchergebnissen zufrieden.	ja	72	R	-	0,34	0,56	12,40	**0,00045**[a]	0,50	0,60	I_{SE}	0,055	0,81
												I_{SP}	0,38	0,68
												I_{EP}	0,04	0,96
												I_{SEP}	0,49	0,61
F23	Ich bin mit meiner Suchleistung zufrieden.	ja	72	R	-	0,12	0,73	7,19	**0,0074**[a]	0,60	0,55	I_{SE}	0,043	0,84
												I_{SP}	0,24	0,79
												I_{EP}	0,32	0,72
												I_{SEP}	0,36	0,70
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	ja	72	R	-	0,75	0,39	12,50	**0,00043**[a]	0,54	0,58	I_{SE}	0,15	0,69
												I_{SP}	0,41	0,66
												I_{EP}	2,36	0,095
												I_{SEP}	0,0086	0,99

[a] Effekt ist in der Tendenz signifikant.

Fortsetzung auf nächster Seite

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Tab. E.42 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	ja	72	R	-	0,84	0,36	18,32	$\mathbf{2{\cdot}10^{-5}}$	0,65	0,52	I_{SE}	3,00	0,084
												I_{SP}	0,89	0,41
												I_{EP}	0,24	0,79
												I_{SEP}	0,44	0,64
SK-A	Accuracy (EUCS)	ja	72	R	-	2,87	0,091	19,75	$\mathbf{9{,}8{\cdot}10^{-6}}$	0,19	0,82	I_{SE}	0,40	0,53
												I_{SP}	0,61	0,54
												I_{EP}	0,45	0,64
												I_{SEP}	0,33	0,72
SK-C	Content (EUCS)	ja	72	R	-	0,77	0,38	16,57	$\mathbf{5{,}1{\cdot}10^{-5}}$	0,22	0,80	I_{SE}	0,031	0,86
												I_{SP}	0,091	0,91
												I_{EP}	0,16	0,85
												I_{SEP}	0,11	0,89
SK-E-09	EUCS-Skala-2009	ja	72	R	-	0,78	0,38	25,29	$\mathbf{5{,}8{\cdot}10^{-7}}$	0,19	0,83	I_{SE}	0,36	0,55
												I_{SP}	0,39	0,68
												I_{EP}	0,53	0,59
												I_{SEP}	0,35	0,70
SK-E-13	EUCS-Skala-2013	ja	72	K[b]	1/68	2,05	0,16	21,19	$\mathbf{1{,}9{\cdot}10^{-5}}$	1,69	0,19	I_{SE}	0,0024	0,96
												I_{SP}	1,03	0,36
												I_{EP}	1,56	0,21
												I_{SEP}	0,045	0,96
SK-E-88	EUCS-Skala-1988	ja	72	R	-	0,76	0,38	14,33	**0,00016**	1,78	0,17	I_{SE}	0,51	0,47
												I_{SP}	0,41	0,67
												I_{EP}	0,75	0,47
												I_{SEP}	0,53	0,59
SK-G-13	Gesamtskala-2013	ja	72	R	-	2,04	0,15	17,85	$\mathbf{2{,}6{\cdot}10^{-5}}$	0,54	0,58	I_{SE}	0,86	0,35
												I_{SP}	0,91	0,40
												I_{EP}	0,00029	1,00
												I_{SEP}	0,53	0,59

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.42 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
SK-K	Kriteriumsskala	ja	72	R	-	0,46	0,50	24,71	$\mathbf{7{,}8{\cdot}10^{-7}}$	3,46	0,032	I_{SE}	1,15	0,28
												I_{SP}	0,0016	1,00
												I_{EP}	0,71	0,49
												I_{SEP}	0,32	0,73
SK-Z-13	Zusatzskala-2013	ja	72	R	-	0,049	0,82	18,34	$\mathbf{2{\cdot}10^{-5}}$	0,71	0,49	I_{SE}	2,74	0,098
												I_{SP}	0,018	0,98
												I_{EP}	0,21	0,81
												I_{SEP}	0,067	0,94
SK01-M	Genauigkeit	ja	72	R	-	0,07	0,79	18,84	$\mathbf{1{,}6{\cdot}10^{-5}}$	0,34	0,71	I_{SE}	0,025	0,87
												I_{SP}	0,062	0,94
												I_{EP}	1,28	0,28
												I_{SEP}	0,30	0,74
SK02-M	Inhalt	ja	72	R	-	2,15	0,14	16,04	$\mathbf{6{,}6{\cdot}10^{-5}}$[a]	1,10	0,33	I_{SE}	0,12	0,73
												I_{SP}	0,96	0,38
												I_{EP}	0,058	0,94
												I_{SEP}	0,27	0,77
SK03-M	Benutzerfreundlichkeit	ja	72	R	-	4,79	0,029	18,39	$\mathbf{2{\cdot}10^{-5}}$[a]	2,90	0,055	I_{SE}	5,18	0,023
												I_{SP}	0,57	0,57
												I_{EP}	1,59	0,21
												I_{SEP}	0,26	0,77
SK04-M	Suche	ja	72	R[c]	-	0,49	0,48	11,60	$\mathbf{0{,}00069}$[a]	1,22	0,29	I_{SE}	0,0025	0,96
												I_{SP}	0,65	0,52
												I_{EP}	0,04	0,96
												I_{SEP}	0,24	0,78
SK08-M	Suche	ja	72	R	-	0,87	0,35	13,96	$\mathbf{2{\cdot}10^{-4}}$	0,66	0,52	I_{SE}	0,00074	0,98
												I_{SP}	0,37	0,69
												I_{EP}	0,074	0,93
												I_{SEP}	0,60	0,55

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.42 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
SK09-M	Benutzerfreundlichkeit	ja	72	R	-	8,73	**0,0032**[a]	13,42	**0,00026**	0,47	0,63	I_{SE}	4,28	0,039
												I_{SP}	1,96	0,14
												I_{EP}	2,87	0,057
												I_{SEP}	0,048	0,95
SK09-M	Benutzerfreundlichkeit	ja	72	R	-	6,77	**0,0094**[a]	17,28	$\mathbf{3{,}5{\cdot}10^{-5}}$	1,24	0,29	I_{SE}	3,62	0,057
												I_{SP}	0,73	0,48
												I_{EP}	1,25	0,29
												I_{SEP}	0,37	0,69
SK11-M[d]		nein	116	R	-	0,73	0,39	13,18	$\mathbf{3{\cdot}10^{-4}}$	0,18	0,83	I_{SE}	0,54	0,46
												I_{SP}	2,56	0,078
												I_{EP}	0,27	0,76
												I_{SEP}	1,39	0,25
SK12-M	Suchergebnis	ja	72	R[c]	-	3,90	0,048	10,37	**0,0013**	0,58	0,56	I_{SE}	5,01	0,025
												I_{SP}	0,54	0,58
												I_{EP}	0,76	0,47
												I_{SEP}	0,88	0,42
SK-E[e]		nein	116	R	-	1,73	0,19	19,63	$\mathbf{1{\cdot}10^{-5}}$	3,47	0,031	I_{SE}	1,92	0,17
												I_{SP}	1,06	0,35
												I_{EP}	2,01	0,13
												I_{SEP}	0,22	0,81
SK14-M	Suche	ja	72	R	-	2,17	0,14	14,06	**0,00019**	0,71	0,49	I_{SE}	0,48	0,49
												I_{SP}	0,17	0,85
												I_{EP}	0,60	0,55
												I_{SEP}	0,029	0,97
SK16-M	Aufgabe	ja	72	R	-	1,21	0,27	15,76	$\mathbf{7{,}7{\cdot}10^{-5}}$	2,10	0,12	I_{SE}	0,28	0,60
												I_{SP}	0,65	0,52
												I_{EP}	0,082	0,92
												I_{SEP}	0,71	0,49

[a] Effekt ist in der Tendenz signifikant.

[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.

[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.42 (Fortsetzung)

						System		Erwartung		Position		Interaktion		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**I**	**F**	**p**
SK17-M	Suche	ja	72	R	-	1,41	0,24	20,53	**$6{,}6{\cdot}10^{-6}$**	0,19	0,83	I_{SE}	0,008	0,93
												I_{SP}	0,27	0,77
												I_{EP}	0,48	0,62
												I_{SEP}	0,45	0,64
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	ja	72	R	-	2,28	0,13	19,32	**$1{,}2{\cdot}10^{-5}$**	1,31	0,27	I_{SE}	0,17	0,68
												I_{SP}	0,65	0,52
												I_{EP}	1,13	0,32
												I_{SEP}	0,99	0,37
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	nein	116	R	-	3,15	0,076	9,00	**0,0028**	0,037	0,96	I_{SE}	1,35	0,24
												I_{SP}	0,30	0,74
												I_{EP}	0,91	0,40
												I_{SEP}	0,034	0,97
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	ja	72	R	-	2,09	0,15	21,69	**$3{,}6{\cdot}10^{-6}$**	0,092	0,91	I_{SE}	0,41	0,52
												I_{SP}	0,48	0,62
												I_{EP}	0,79	0,45
												I_{SEP}	0,30	0,74
E06-M	Erwartungsskala	ja	72	R	-	7,37	0,0067	11,18	**0,00086**	0,56	0,57	I_{SE}	2,13	0,14
												I_{SP}	0,72	0,49
												I_{EP}	0,55	0,58
												I_{SEP}	0,63	0,53

[a] Effekt ist in der Tendenz signifikant.
[b] Für einzelne Stichproben wird eine robuste Varianzanalyse durchgeführt.
[c] Für einzelne Stichproben wird eine klassische Varianzanalyse durchgeführt.
[d] Entspricht auch den Skalen SK15-M und SK19-M.
[e] Entspricht auch der Skala SK13-M.

Tab. E.43.: Teststatistik der Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerzufriedenheit in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

						System		**Erwartung**		**Position**		**Interaktion**		
ID	**Beschreibung**	**topic**	**n**	**V**	**df**	**F**	**p**	**F**	**p**	**F**	**p**	**Art**	**F**	**p**
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	anz	80	R	-	0,32	0,57	23,34	$\mathbf{1{,}6{\cdot}10^{-6}}$	0,68	0,51	I_{SE}	0,32	0,57
												I_{SP}	0,7	0,50
												I_{EP}	0,097	0,91
												I_{SEP}	1,12	0,33
F12	Ist die Suchmaschine erfolgreich?	anz	80	R	-	0,41	0,52	19,01	$\mathbf{1{,}4{\cdot}10^{-5}}$	0,89	0,41	I_{SE}	0,62	0,43
												I_{SP}	0,014	0,99
												I_{EP}	0,19	0,82
												I_{SEP}	0,066	0,94
SK11-M[a]	Eigenleistung	anz	80	R	-	0,19	0,66	19,47	$\mathbf{1{,}1{\cdot}10^{-5}}$	0,65	0,52	I_{SE}	0,084	0,77
												I_{SP}	0,47	0,62
												I_{EP}	0,14	0,87
												I_{SEP}	0,81	0,45
SK-E[b]	Ease of Use (EUCS)	anz	80	R	-	1,23	0,27	22,73	$\mathbf{2{,}1{\cdot}10^{-6}}$	0,98	0,37	I_{SE}	1,69	0,19
												I_{SP}	0,13	0,87
												I_{EP}	1,55	0,21
												I_{SEP}	0,28	0,75
E04	Wie wahrsch. ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchm. erbringen, sehr überzeugt sind?	anz	80	R	-	2,62	0,11	10,00	**0,0016**	0,038	0,96	I_{SE}	1,00	0,32
												I_{SP}	0,24	0,79
												I_{EP}	0,92	0,40
												I_{SEP}	0,2	0,82

[a] Entspricht auch den Skalen SK15-M und SK19-M.

[b] Entspricht auch der Skala SK13-M.

Tab. E.44.: Teststatistik signifikanter Interaktionen der klassischen Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerzufriedenheit in SP_A. Fett hervorgehoben sind Effekte, die in mindestens einer von fünf Stichproben nachweisbar sind.

				System			Erwartung			Position			Interaktion			
ID	**Beschreibung**	**topic**	**n**	**df**	**F**	**p**	**df**	**F**	**p**	**df**	**F**	**p**	**Art**	**df**	**F**	**p**
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	ja	72	1/68	6,79	0,011	1/68	10,15	0,0022	2/136	0,055	0,95	I_{SE}	1/68	4,59	**0,036**
													I_{SP}	2/136	0,27	0,76
													I_{EP}	2/136	0,18	0,83
													I_{SEP}	2/136	1,18	0,31
F04	Liefert die Suchmaschine genügend Information?	ja	72	1/68	6,16	0,016	1/68	11,20	0,0013	2/136	2,68	0,072	I_{SE}	1/68	2,62	0,11
													I_{SP}	2/136	3,28	**0,041**
													I_{EP}	2/136	3,59	**0,03**
													I_{SEP}	2/136	1,73	0,18
F08	Ist die Suchmaschine benutzerfreundlich?	ja	72	1/68	0,54	0,46	1/68	11,22	0,0013	2/136	0,24	0,78	I_{SE}	1/68	3,51	0,065
													I_{SP}	2/136	0,55	0,58
													I_{EP}	2/136	4,90	**0,0088**
													I_{SEP}	2/136	0,027	0,97
				1/68	0,61	0,44	1/68	7,82	0,0067	2/136	0,94	0,39	I_{SE}	1/68	1,25	0,27
													I_{SP}	2/136	3,79	**0,025**
													I_{EP}	2/136	1,46	0,24
													I_{SEP}	2/136	0,37	0,69
F13	Sind Sie mit der Suchmaschine zufrieden?	ja	72	1/68	0,056	0,81	1/68	6,33	0,014	2/136	5,33	0,0059	I_{SE}	1/68	0,18	0,67
													I_{SP}	2/136	0,52	0,60
													I_{EP}	2/136	4,65	**0,011**
													I_{SEP}	2/136	0,30	0,74
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	ja	72	1/68	0,93	0,34	1/68	2,15	0,15	2/136	1,49	0,23	I_{SE}	1/68	3,56	0,064
													I_{SP}	2/136	4,25	**0,016**
													I_{EP}	2/136	1,96	0,15
													I_{SEP}	2/136	1,12	0,33

[a] Entspricht auch der Skala SK18-M.

[b] Entspricht auch der Skala SK13-M.

[c] Dieser Effekt ist in zwei Zufallsstichproben signifikant.

Fortsetzung auf nächster Seite

Tab. E.44 (Fortsetzung)

				System			Erwartung			Position			Interaktion			
ID	**Beschreibung**	**topic**	**n**	**df**	**F**	**p**	**df**	**F**	**p**	**df**	**F**	**p**	**I**	**df**	**F**	**p**
F23	Ich bin mit meiner Suchleistung zufrieden.	ja	72	1/68	0,54	0,46	1/68	6,76	0,011	2/136	0,95	0,39	I_{SE}	1/68	1,27	0,26
													I_{SP}	2/136	3,86	**0,023**
													I_{EP}	2/136	0,43	0,65
													I_{SEP}	2/136	2,19	0,12
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	ja	72	1/68	0,014	0,90	1/68	19,66	$3{,}5{\cdot}10^{-5}$	2/136	1,71	0,19	I_{SE}	1/68	0,52	0,47
													I_{SP}	2/136	1,08	0,34
													I_{EP}	2/136	5,69	**0,0042**
													I_{SEP}	2/136	0,46	0,63
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	ja	72	1/68	0,40	0,53	1/68	16,76	0,00011	2/136	1,68	0,19	I_{SE}	1/68	2,44	0,12
													I_{SP}	2/136	3,27	**0,041**
													I_{EP}	2/136	0,56	0,57
													I_{SEP}	2/136	0,83	0,44
SK03-M	Benutzerfreundlichkeit	ja	72	1/68	2,47	0,12	1/68	19,09	$4{,}4{\cdot}10^{-5}$	2/136	2,54	0,083	I_{SE}	1/68	6,57	**0,013**[c]
													I_{SP}	2/136	1,88	0,16
													I_{EP}	2/136	0,65	0,53
													I_{SEP}	2/136	1,06	0,35
SK07-M[a]	Benutzerfreundlichkeit	ja	72	1/68	2,01	0,16	1/68	4,07	0,048	2/136	1,47	0,23	I_{SE}	1/68	0,074	0,79
													I_{SP}	2/136	0,37	0,69
													I_{EP}	2/136	3,79	**0,025**[c]
													I_{SEP}	2/136	0,13	0,87
SK09-M	Benutzerfreundlichkeit	ja	72	1/68	2,49	0,12	1/68	9,36	0,0032	2/136	1,94	0,15	I_{SE}	1/68	12,17	**0,00086**[c]
													I_{SP}	2/136	2,28	0,11
													I_{EP}	2/136	1,96	0,15
													I_{SEP}	2/136	0,018	0,98
				1/68	6,66	0,012	1/68	13,12	0,00056	2/136	1,58	0,21	I_{SE}	1/68	4,53	0,037
													I_{SP}	2/136	1,95	0,15
													I_{EP}	2/136	4,44	**0,014**
													I_{SEP}	2/136	0,037	0,96

[a] Entspricht auch der Skala SK18-M.

[b] Entspricht auch der Skala SK13-M.

[c] Dieser Effekt ist in zwei Zufallsstichproben signifikant.

Fortsetzung auf nächster Seite

Tab. E.44 (Fortsetzung)

				System			Erwartung			Position			Interaktion			
ID	**Beschreibung**	**topic**	**n**	**df**	**F**	**p**	**df**	**F**	**p**	**df**	**F**	**p**	**I**	**df**	**F**	**p**
SK12-M	Suchergebnis	ja	72	1/68	2,22	0,14	1/68	12,36	0,00079	2/136	1,71	0,19	I_{SE}	1/68	5,89	**0,018**
													I_{SP}	2/136	0,63	0,53
													I_{EP}	2/136	1,08	0,34
													I_{SEP}	2/136	1,59	0,21
SK-E[b]	Ease of Use (EUCS)	nein	116	1/112	0,46	0,50	1/112	21,27	$1{,}1{\cdot}10^{-5}$	2/224	8,02	0,00043	I_{SE}	1/112	1,78	0,18
													I_{SP}	2/224	1,32	0,27
													I_{EP}	2/224	3,08	**0,048**
													I_{SEP}	2/224	0,37	0,69
SK-T	Timeliness (EUCS)	ja	72	1/68	0,54	0,46	1/68	9,57	0,0029	2/136	5,64	0,0044	I_{SE}	1/68	0,73	0,40
													I_{SP}	2/136	2,33	0,10
													I_{EP}	2/136	8,44	**0,00035**
													I_{SEP}	2/136	0,41	0,67
SK-G-13	Gesamtskala-2013	ja	72	1/68	3,85	0,054	1/68	7,87	0,0065	2/136	1,53	0,22	I_{SE}	1/68	4,45	**0,039**
													I_{SP}	2/136	1,24	0,29
													I_{EP}	2/136	1,31	0,27
													I_{SEP}	2/136	1,85	0,16
E03	Wie wahrsch. ist es, dass Sie mithilfe dieser Suchm. zu einem schnellen Ergebnis kommen?	ja	72	1/68	0,31	0,58	1/68	16,52	0,00013	2/136	5,11	0,0072	I_{SE}	1/68	0,14	0,71
													I_{SP}	2/136	0,088	0,92
													I_{EP}	2/136	4,32	**0,015**
													I_{SEP}	2/136	1,87	0,16
E06-M	Erwartungsskala	ja	72	1/68	0,51	0,48	1/68	5,80	0,019	2/136	2,19	0,12	I_{SE}	1/68	0,47	0,50
													I_{SP}	2/136	0,15	0,86
													I_{EP}	2/136	3,29	**0,04**
													I_{SEP}	2/136	0,27	0,76

[a] Entspricht auch der Skala SK18-M.
[b] Entspricht auch der Skala SK13-M.
[c] Dieser Effekt ist in zwei Zufallsstichproben signifikant.

Tab. E.45.: Teststatistik signifikanter Interaktionen der klassischen Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerzufriedenheit in SP_B. Fett hervorgehoben sind Effekte, die in mindestens einer von fünf Stichproben nachweisbar sind.

				System			Erwartung			Position			Interaktion			
ID	**Beschreibung**	**topic**	**n**	**df**	**F**	**p**	**df**	**F**	**p**	**df**	**F**	**p**	**Art**	**df**	**F**	**p**
SK-E[a]	Ease of Use (EUCS)	nein	40	1/76	0,27	0,60	1/76	22,55	$9{,}4 \cdot 10^{-06}$	2/152	5,56	0,0047	I_{SE}	1/76	1,15	0,29
													I_{SP}	2/152	0,74	0,48
													I_{EP}	2/152	3,71	**0,027**[b]
													I_{SEP}	2/152	0,94	0,39

[a] Entspricht auch der Skala SK13-M.

[b] Dieser Effekt ist in drei Zufallsstichproben signifikant.

Tab. E.46.: Teststatistik signifikanter Interaktionen der klassischen Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung in SP_A. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

				System			Erwartung			Position			Interaktion			
ID	**Beschreibung**	**topic**	**n**	**df**	**F**	**p**	**df**	**F**	**p**	**df**	**F**	**p**	**Art**	**df**	**F**	**p**
M27[a]	Anz. falsch eher irrel. bew. eher rel. Dok.	nein	116	1/112	4,67	**0,033**[b]	1/112	0,46	0,50	1/112	3,50	**0,032**[b]	I_{SE}	1/112	1,17	0,28
													I_{SP}	2/224	0,73	0,48
													I_{EP}	2/224	0,73	0,48
													I_{SEP}	2/224	3,71	**0,026**
B05[a]	Durchschn. Bew. rel. Dok. (erste Suche)	nein	72	1/68	0,35	0,56	1/68	0,27	0,61	1/68	0,021	0,98	I_{SE}	1/68	0,45	0,50
													I_{SP}	2/136	4,88	**0,009**
													I_{EP}	2/136	0,045	0,96
													I_{SEP}	2/136	0,66	0,52
Z07[a]	Durchschn. Betrachtungsz. rel. bew. Dok.	ja	72	1/68	0,0068	0,93	1/68	0,00085	0,98	1/68	5,85	**0,0096**	I_{SE}	1/68	7,73	**0,007**
													I_{SP}	2/136	0,52	0,53
													I_{EP}	2/136	0,21	0,73
													I_{SEP}	2/136	0,99	0,35

[a] Normalverteilungsvoraussetzung verletzt.

[b] In der Tendenz signifikant.

Fortsetzung auf nächster Seite

Tab. E.46 (Fortsetzung)

				System			Erwartung			Position			Interaktion			
ID	**Beschreibung**	**topic**	**n**	**df**	**F**	**p**	**df**	**F**	**p**	**df**	**F**	**p**	**I**	**df**	**F**	**p**
S04[a]	Suchdauer	nein	116	1/112	0,02	0,90	1/112	0,0035	0,95	1/112	13,95	$\mathbf{1{,}9{\cdot}10^{-6}}$	I_{SE}	1/112	0,19	0,67
													I_{SP}	2/224	4,84	**0,0088**
													I_{EP}	2/224	0,70	0,50
													I_{SEP}	2/224	2,70	0,070

[a] Normalverteilungsvoraussetzung verletzt.
[b] In der Tendenz signifikant.

Tab. E.47.: Teststatistik signifikanter Interaktionen der klassischen Varianzanalyse zur Untersuchung des dynamischen Einflusses von System und Erwartung auf die Benutzerleistung in SP_B. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind.

				System			Erwartung			Position			Interaktion			
ID	**Beschreibung**	**topic**	**n**	**df**	**F**	**p**	**df**	**F**	**p**	**df**	**F**	**p**	**Art**	**df**	**F**	**p**
M27[a]	Anz. falsch eher irrel. bew. eher rel. Dok.	nein	80	1/76	3,18	0,079	1/76	0,73	0,40	1/76	1,47	0,23	I_{SE}	1/76	4,38	0,04
													I_{SP}	2/152	0,74	0,48
													I_{EP}	2/152	0,59	0,55
													I_{SEP}	2/152	4,89	**0,0087**
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	80	1/76	0,014	0,91	1/76	3,46	0,067[c]	1/76	4,92	**0,0085**	I_{SE}	1/76	0,41	0,52
													I_{SP}	2/152	3,90	**0,036**
													I_{EP}	2/152	1,42	0,25
													I_{SEP}	2/152	1,52	0,22
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	80	1/76	0,0076	0,93	1/76	4,69	**0,033**	1/76	4,71	**0,010**	I_{SE}	1/76	0,94	0,34
													I_{SP}	2/152	3,56	**0,031**[b]
													I_{EP}	2/152	1,15	0,32
													I_{SEP}	2/152	1,89	0,15
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	nein	72	1/68	0,81	0,37	1/68	3,54	0,064	1/68	3,04	0,051	I_{SE}	1/68	0,59	0,45
													I_{SP}	2/136	5,03	**0,0078**
													I_{EP}	2/136	1,71	0,19
													I_{SEP}	2/136	0,68	0,51

[a] Normalverteilungsvoraussetzung verletzt.
[b] In der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.

E.5. Weitere Ergebnisse der Topiceffektanalyse

In diesem Abschnitt sind weitergehende Ergebnisse der Topiceffektanalyse von Experiment 3 zusammengefasst. Dabei gliedert sich der Abschnitt in drei Teile. In Abschnitt E.5.1 findet sich zunächst erneut eine Übersicht derjenigen Variablen, für die sich in keiner der untersuchten Stichproben und bei keiner der drei Testaufgaben ein signifikanter Topiceffekt nachweisen lässt. Abschnitt E.5.2 fasst die Ergebnisse der im dritten Experiment aufgrund der Problematik im Zusammenhang mit der Formulierung der Suchanfragen für das Wikithema notwendigen Analyse orthogonaler Kontraste zusammen. Die Ergebnisse der Post-hoc-Tests unter Ausschluss kritischer Fallgruppen können hingegen in Abschnitt E.5.3 nachvollzogen werden.

E.5.1. Variablen ohne signifikante Unterschiede

Wie auch schon im zweiten Experiment lässt sich für einige Variablen in keiner der untersuchten Stichproben und bei keiner der Testaufgaben ein signifikanter Topiceffekt feststellen. Eine Übersicht der betreffenden Variablen zeigt Tabelle E.48. Neben der Stichprobengröße gibt die Tabelle auch Auskunft darüber, ob eine klassische (K) oder robuste (R) Varianzanalyse durchgeführt wird. Eine Fußnote markiert darüber hinaus Fälle, in denen eine der beiden Analysevarianten die Ausnahme bleibt.

Tab. E.48.: Übersicht der abhängigen Variablen, die keinen Topiceffekt aufweisen. Je nach Gegebenheit der Verteilungsvoraussetzungen wurde über alle fünf Stichproben hinweg oder einzeln pro Stichprobe eine klassische (K) oder robuste (R) Varianzanalyse durchgeführt.

		SP_A		SP_B	
ID	**Beschreibung**	**n**	**Anova**	**n**	**Anova**
M01	Anz. aufg. Dok.	72	R	24	K/R[a]
M05	Anz. aufg. irrel. Dok.	72	R	24	R
M07	Anz. falsch irrel. bew. Dok.	72	R	24	R
M08	Anz. falsch rel. bew. Dok.	72	R	24	R
M10	Anz. rel. bew. Dok.	72	R	24	R
M13	Anz. rel. bew. Dok. (letzte Suche)	72	R	24	R
M18	Anz. richtig rel. bew. Dok. (erste Suche)	72	R	24	R
M20	Anz. aufg. eher irrel. Dok.	72	R	24	R
M25	Anz. eher rel. bew. Dok.	72	R	24	R
M26	Anz. falsch eher irrel. bew. Dok.	72	R	24	R
M27	Anz. falsch eher irrel. bew. eher rel. Dok.	72	R	24	R
M28	Anz. falsch eher irrel. bew. irrel. Dok.	72	R	24	R
M29	Anz. falsch eher irrel. bew. rel. Dok.	72	R	24	R
M31	Anz. falsch eher rel. bew. eher irrel. Dok.	72	R	24	R
M32	Anz. falsch eher rel. bew. irrel. Dok.	72	R	24	R
M37	Anz. rel. bew. Dok. (4-st.)	72	R	24	R
M40	Anz. rel. bew. Dok. (letzte Suche) (4-st.)	72	R	24	R
M41	Anz. richtig bew. Dok. (4-st.)	72	R	24	R
M42	Anz. richtig eher irrel. bew. Dok.	72	R	24	R
M46	Anz. richtig rel. bew. Dok. (erste 10 Dok.) (4-st.)	72	R	24	R

[a] Nur einmal klassische Analyse.

Fortsetzung auf nächster Seite

[b] Nur einmal robuste Anaylse.

[c] Entspricht auch den Skalen SK15-M und SK19-M.

[d] Entspricht auch der Skala SK13-M.

Tab. E.48 (Fortsetzung)

		SP_A		SP_B	
ID	**Beschreibung**	**n**	**Anova**	**n**	**Anova**
M48	Anz. richtig rel. bew. Dok. (letzte Suche) (4-st.)	72	R	24	R
B04	Durchschn. Bew. rel. Dok.	72	K/R[a]	24	K/R[b]
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	48	R	24	K/R[b]
B16	Durchschn. Bew. rel. Dok. (4-st.)	72	R	24	K/R[a]
Z01-log	Durchschn. Betrachtungsz. aller Dok.	72	K	24	K/R
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	72	K/R	24	K
Z08	Durchschn. Betrachtungsz. rel. Dok.	72	R	24	R
Z08-log	Durchschn. Betrachtungsz. rel. Dok.	72	K/R[b]	24	K
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	72	R	24	K/R[a]
Z11	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	72	R	24	R
Z11-log	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	72	K/R[b]	24	K
Z23	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	72	R	24	R
Z24	Durchschn. Betrachtungsz. richtig bew. Dok. (4-st.)	48	R	24	R
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	72	R	24	K/R
V02	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	72	R	24	K/R
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	72	R	24	R
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	72	R	24	R
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	72	K/R[a]	24	K/R
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	72	K/R[a]	24	K/R[b]
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	72	R	24	R
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	72	K/R	24	K/R
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	72	R	24	K/R[a]
V31/BP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	72	R	24	R
V32/BR	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	72	R	24	K/R[a]
V33	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	72	R	24	K/R[a]

[a] Nur einmal klassische Analyse.

[b] Nur einmal robuste Anaylse.

[c] Entspricht auch den Skalen SK15-M und SK19-M.

[d] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.48 (Fortsetzung)

		SP_A		SP_B	
ID	**Beschreibung**	**n**	**Anova**	**n**	**Anova**
V55	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	R	24	K/R
V56	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	R	24	R
V61	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	R	24	R
S02	Erste betr. Rankingpos.	72	R	24	R
S03	Letzte betr. Rankingpos.	72	R	24	R
S04	Suchdauer	72	R	24	R
S05-log	Zeit zum ersten richtig rel. bew. Dok.	48	K	24	K/R[b]
F11	Liefert die Suchmaschine aktuelle Information?	72	R	24	R
F12	Ist die Suchmaschine erfolgreich?	72	R	24	R
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	72	R	24	R
SK11-M[c]	Eigenleistung	72	R	24	K/R
SK-E[d]	Ease of Use (EUCS)	72	R	24	R
E01	Ich glaube, ich werde in zehn Minuten ... rel. Dok. finden.	72	R	24	K/R[a]
E04	Wie wahrsch. ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchm. erbringen, sehr überzeugt sind?	72	R	24	R

[a] Nur einmal klassische Analyse.
[b] Nur einmal robuste Analyse.
[c] Entspricht auch den Skalen SK15-M und SK19-M.
[d] Entspricht auch der Skala SK13-M.

E.5.2. Orthogonale Kontraste

Zusätzlich zu den in Abschnitt 7.4.6.1 beschriebenen Post-Hoc-Tests, wird im Rahmen der Topiceffektanalyse noch einmal explizit der Unterschied zwischen dem Wiki- und den anderen beiden Suchthemen analysiert. Zu diesem Zweck werden zwei orthogonale Kontraste für einen weiteren Post-Hoc-Test definiert: Der erste Kontrast (Wiki-Rest) prüft, ob das Wikithema im Durchschnitt schwieriger ist, als die anderen beiden Suchthemen. Der zweite Kontrast (Wind-Englisch) prüft den Unterschied zwischen dem Wind- und dem Englischthema. Die entsprechenden Ergebnisse sind den Tabellen E.49 und E.50 zu entnehmen. Da dieser zusätzliche Test ohne robuste Methoden durchgeführt wird, wird hier unabhängig von der Frage, ob die Verteilungsvoraussetzungen im Einzelfall erfüllt sind, das signifikanteste Ergebnis der Varianzanalyse mit Messwiederholung berichtet. Aus diesem Grund sind die folgenden Ergebnisse als explorative und ergänzende Analyse zu bewerten. Zu beachten ist weiterhin, dass sich nicht in jedem Fall für alle fünf Stichproben ein signifikanter Aufgabeneffekt nachweisen lässt. Die entsprechenden Fälle sind in den Tabellen durch eine Fußnote gekennzeichnet. Umgekehrt kann es auch passieren, dass Kontraste zwar für einzelne Stichproben, aber nicht für die Varianzanalyse mit dem niedrigsten p-Wert signifikant werden. Auch diese Fälle sind in den Tabellen kenntlich gemacht.

Tab. E.49.: Signifikante Ergebnisse der klassischen Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung mit Hilfe orthogonaler Kontraste in SP_A. Die Auswertung der orthogonalen Kontraste erfolgt unabhängig davon, ob die Verteilungsvoraussetzungen zur Durchführung einer klassischen Varianzanalyse erfüllt sind. Darüber hinaus werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	df	t
M14	Anz. richtig bew. Dok.	72	**0,0233**	Wiki–Rest[a]	-0,36	142	-2,29
			0,0881	Wind–Englisch	-0,47	142	-1,72
M35	Anz. falsch rel. bew. Dok. (4-st.)	72	0,067[d]	Wiki–Rest	-0,09	142	-1,85
			$2{\cdot}10^{-4}$	Wind–Englisch[a]	0,31	142	3,89
M36	Anz. irrel. bew. Dok. (4-st.)	72	0,7876	Wiki–Rest	-0,02	142	-0,27
			0,0049[b]	Wind[a]Englisch	-0,38	142	-2,86
B10	Durchschn. Bew. eher rel. Dok.	48	0,6278	Wiki–Rest	-0,04	94	-0,49
			0,0011	Wind[a]Englisch	0,50	94	3,38
V09	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	72	0,879	Wiki–Rest	0,001	142	0,15
			0,0087	Wind–Englisch[a]	0,03	142	2,66
V34	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$	72	**$4{\cdot}10^{-4}$**	Wiki–Rest[a]	-0,01	142	-3,63
			0,0256[b]	Wind[a]Englisch	0,01	142	2,26
V35	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	72	**0,0036**	Wiki–Rest[a]	-0,01	142	-2,96
			0,0049	Wind[a]Englisch	0,01	142	2,86
V37	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	0,0933[d]	Wiki–Rest	0,01	142	1,69
			$1{\cdot}10^{-4}$	Wind–Englisch[a]	-0,06	142	-3,93
V38	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$ (4-st.)	72	0,8412[d]	Wiki–Rest	-0,0004	142	-0,20
			0,0039	Wind–Englisch[a]	-0,01	142	-2,93
V51	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	0,1579[d]	Wiki–Rest	-0,01	142	-1,42
			$7{\cdot}10^{-4}$	Wind–Englisch[a]	0,04	142	3,48
V52	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$ (4-st.)	24	0,0029[b]	Wiki[a]Rest	-0,07	46	-3,14
			0,0022[b]	Wind–Englisch[a]	0,12	46	3,24
V54	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	0,1595	Wiki–Rest	-0,01	142	-1,41
			0,0308[b]	Wind[a]Englisch	-0,03	142	-2,18
V79	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$ (4-st.)	24	0,0937[d]	Wiki–Rest	0,01	46	1,71
			0,0265[b]	Wind–Englisch[a]	-0,01	46	-2,29
S01	Anz. Suchen	72	0,789	Wiki–Rest	-0,02	142	-0,27
			0,0217[b]	Wind–Englisch	-0,24	142	-2,32

[a] Dieses Thema entspricht der leichteren Aufgabe.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.

Tab. E.50.: Signifikante Ergebnisse der klassischen Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit mit Hilfe orthogonaler Kontraste in SP_A. Die Auswertung der orthogonalen Kontraste erfolgt unabhängig davon, ob die Verteilungsvoraussetzungen zur Durchführung einer klassischen Varianzanalyse erfüllt sind. Darüber hinaus werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	df	t
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	72	0,4063	Wiki–Rest	0,04	142	0,83
			$1{\cdot}10^{-4}$	Wind[a]Englisch	0,31	142	3,97
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	72	0,743	Wiki–Rest	-0,01	142	-0,33
			$1{\cdot}10^{-4}$	Wind[a]Englisch	0,29	142	3,98
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	72	0,0167[b]	Wiki–Rest[a]	-0,11	142	-2,42
			0	Wind[a]Englisch	0,34	142	4,20
F22	Ich bin mit den Suchergebnissen zufrieden.	72	0,3816	Wiki–Rest	-0,04	142	-0,88
			0,0028	Wind[a]Englisch	0,25	142	3,04
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	72	0,1002[d]	Wiki–Rest	-0,07	142	-1,65
			0,0137[b]	Wind[a]Englisch	0,19	142	2,50
SK02-M	Inhalt	72	0,9433	Wiki–Rest	-0,003	142	-0,07
			0,0017	Wind[a]Englisch	0,24	142	3,21
SK04-M	Suche	72	0,1473	Wiki–Rest	-0,06	142	-1,46
			0,001	Wind[a]Englisch	0,24	142	3,35
SK08-M	Suche	72	0,1637	Wiki–Rest	-0,06	142	-1,40
			$7{\cdot}10^{-4}$	Wind[a]Englisch	0,24	142	3,47
SK-E-09	EUCS-Skala-2009	72	0,3098	Wiki–Rest	0,03	142	1,02
			0,0015	Wind[a]Englisch	0,19	142	3,24
SK14-M	Suche	72	0,8573	Wiki–Rest	-0,01	142	-0,18
			$1{\cdot}10^{-4}$	Wind[a]Englisch	0,27	142	4,06
SK16-M	Aufgabe	72	0,0252[b]	Wiki–Rest[a]	-0,10	142	-2,26
			$1{\cdot}10^{-4}$	Wind[a]Englisch	0,30	142	4,01
SK17-M	Suche	72	0,3053	Wiki–Rest	-0,04	142	-1,03
			0,0021	Wind[a]Englisch	0,21	142	3,13
SK-A	Accuracy (EUCS)	72	0,4251	Wiki–Rest	-0,03	142	-0,80
			0,0023	Wind[a]Englisch	0,23	142	3,11
SK-E-13	EUCS-Skala-2013	72	0,156	Wiki–Rest	0,05	142	1,43
			0,0021	Wind[a]Englisch	0,17	142	3,14
SK-Z-13	Zusatzskala-2013	72	0,5724	Wiki–Rest	-0,02	142	-0,57
			0,0014	Wind[a]Englisch	0,21	142	3,26
SK-G-13	Gesamtskala-2013	72	0,9705	Wiki–Rest	-0,001	142	-0,04
			$2{\cdot}10^{-4}$	Wind[a]Englisch	0,24	142	3,86

[a] Bei diesem Thema sind die Probanden zufriedener.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.

E.5.3. Ergebnisse der Topiceffektanalyse unter Ausschluss kritischer Fallgruppen

In diesem Abschnitt sind die Ergebnisse der Post-hoc-Tests für die unter Ausschluss bestimmter als problematisch eingeschätzter Fallgruppen durchgeführten Topiceffektanalysen zur Verfügung gestellt. Wie bei der Auswertung der Gesamtdaten wird auch an dieser Stelle, aufgrund der geringen Anzahl an Variablen für die die Verteilungsvoraussetzungen erfüllt sind, auf die Nennung der klassischen Analyseergebnisse verzichtet. Der Aufbau der Tabellen ist identisch mit dem der in Abschnitt 7.4.6.1 aufgeführten Tabellen, sodass eine Erklärung der Lesart entfällt. Neben der Frage, ob ein bestimmter Mittelwertunterschied in der Tendenz oder über alle fünf Stichproben hinweg (in den Tabellen fett hervorgehoben) stabil ist, informieren zusätzlich in Fußnoten eingefügte Erläuterungen über weitere Trends in den Gesamtdaten, wie bspw. das zufällige Signifikantwerden einzelner Topicunterschiede oder die Bestätigung der im Rahmen der Hauptanalyse erhaltenen Befunde. Bevor jedoch in den Tabellen E.52 bis E.58 die Ergebnisse der unter Ausschluss problematischer Fallgruppen durchgeführten Post-Hoc-Tests angegeben werden, informiert Tabelle E.51 zunächst noch kurz über im Kontext der Benutzerleistung signifikant werdende Topiceffekte in der Stichprobe SP_B, die aus Platzgründen nicht innerhalb der Arbeit behandelt werden.

Tab. E.51.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung in SP_B. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
M45	Anz. richtig rel. bew. Dok. (4-st.)	24	**$< 0{,}05$**	Wind–Englisch[a]	-1	$-1{,}88$	$-0{,}12$
			$> 0{,}05$	Wind–Wiki	$0{,}12$	$-0{,}93$	$1{,}18$
			$< 0{,}05$[b]	Englisch[a]Wiki	$1{,}12$	$0{,}10$	$2{,}15$
V35[e]	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	24	$> 0{,}05$[d]	Wind–Englisch	$0{,}03$	$-0{,}01$	$0{,}06$
			$> 0{,}05$	Wind–Wiki	$-0{,}02$	$-0{,}06$	$0{,}02$
			$< 0{,}05$[b]	Englisch–Wiki[a]	$-0{,}05$	$-0{,}09$	$-0{,}01$
V37[e]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	24	$> 0{,}05$	Wind–Englisch	$-0{,}18$	$-0{,}37$	$0{,}01$
			$> 0{,}05$	Wind–Wiki	$-0{,}03$	$-0{,}18$	$0{,}11$
			$> 0{,}05$	Englisch–Wiki	$0{,}14$	$-0{,}005$	$0{,}29$

[a] Dieses Thema entspricht der leichteren Aufgabe.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.52.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung in SP_A unter Ausschluss von SP_{TD}. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
M21	Anz. aufg. eher rel. Dok.	72	$> 0{,}05$[d]	Wind–Englisch	0,64	-0,06	1,33
			$> 0{,}05$	Wind–Wiki	-0,45	-1,12	0,21

[a] Dieses Thema entspricht der leichteren Aufgabe.
Fortsetzung auf nächster Seite
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.52 (Fortsetzung)

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
			< 0,05	Englisch–Wiki[a]	-1,09	-1,69	-0,50
			< 0,05	Wind–Englisch[a]	-0,98	-1,62	-0,34
M23	Anz. aufg. rel. Dok. (4-st.)	72	> 0,05	Wind–Wiki	-0,66	-1,41	0,09
			> 0,05	Englisch–Wiki	0,32	-0,50	1,14
			< 0,05	Wind[a]Englisch	-0,59	-1,03	-0,15
M33	Anz. falsch eher rel. bew. rel. Dok.	72	> 0,05	Wind–Wiki	-0,14	-0,58	0,31
			> 0,05	Englisch–Wiki	0,45	-0,03	0,94
			< 0,05	Wind–Englisch[a]	0,52	0,15	0,90
M35[e]	Anz. falsch rel. bew. Dok. (4-st.)	72	> 0,05	Wind–Wiki	-0,14	-0,58	0,31
			< 0,05[b]	Englisch[a]Wiki	-0,66	-1,07	-0,25
			< 0,05[b]	Wind[a]Englisch	-0,77	-1,47	-0,08
M36[e]	Anz. irrel. bew. Dok. (4-st.)	72	**< 0,05**	Wind[a]Wiki	-1,05	-1,67	-0,42
			> 0,05	Englisch–Wiki	-0,27	-1,07	0,52
			< 0,05[b]	Wind[a]Englisch	0,48	0,02	0,93
M38	Anz. rel. bew. Dok. (erste 10 Dok.) (4-st.)	72	> 0,05	Wind–Wiki	-0,23	-0,75	0,30
			< 0,05	Englisch–Wiki[a]	-0,70	-1,17	-0,24
			< 0,05[b]	Wind[a]Englisch	1,53	0,35	2,70
B10[e]	Durchschn. Bew. eher rel. Dok.	24	> 0,05[d]	Wind–Wiki	0,36	-0,65	1,37
			< 0,05[b]	Englisch–Wiki[a]	-1,17	-2,27	-0,07
			> 0,05[d]	Wind–Englisch	0,02	-0,001	0,03
V34[e]	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$	72	> 0,05	Wind–Wiki	-0,01	-0,03	0,01
			< 0,05	Englisch–Wiki[a]	-0,02	-0,04	-0,01
			< 0,05	Wind[a]Englisch	0,03	0,004	0,05
V35[e]	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	72	> 0,05	Wind–Wiki	-0,003	-0,03	0,02
			< 0,05	Englisch–Wiki[a]	-0,03	-0,05	-0,01
			< 0,05	Wind–Englisch[a]	-0,13	-0,22	-0,04
V37[e]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	72	> 0,05	Wind–Wiki	-0,03	-0,10	0,03
			< 0,05[b]	Englisch[a]Wiki	0,10	0,01	0,18
			< 0,05	Wind–Englisch[a]	-0,02	-0,04	-0,01
V38[e]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$ (4-st.)	72	> 0,05	Wind–Wiki	-0,01	-0,03	0,01
			> 0,05	Englisch–Wiki	0,01	-0,01	0,03
			< 0,05	Wind–Englisch[a]	-0,04	-0,06	-0,02
V39	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$ (4-st.)	72	> 0,05	Wind–Wiki	-0,02	-0,05	0,01
			> 0,05	Englisch–Wiki	0,02	-0,005	0,05
			< 0,05	Wind–Englisch[a]	0,10	0,04	0,15
V51[e]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	> 0,05	Wind–Wiki	0,03	-0,02	0,09
			< 0,05	Englisch[a]Wiki	-0,06	-0,11	-0,01
			< 0,05[b]	Wind–Englisch	-0,80	-1,25	-0,34
S01[e]	Anz. Suchen	72	> 0,05	Wind–Wiki	-0,34	-0,80	0,12
			> 0,05	Englisch–Wiki	0,45	-0,05	0,96

[a] Dieses Thema entspricht der leichteren Aufgabe.

[b] Nicht mindestens in der Tendenz signifikant.

[c] Eigentlich in der Tendenz signifikant.

[d] Teilweise auch signifikant.

[e] In SP_A mindestens in der Tendenz signifikant.

[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.53.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit in SP_A unter Ausschluss von SP_{TD}. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	72	**< 0,05**	Wind[a]Englisch	0,70	0,16	1,25
			> 0,05	Wind–Wiki	0,39	-0,17	0,94
			> 0,05	Englisch–Wiki	-0,32	-0,85	0,21
F04	Liefert die Suchmaschine genügend Information?	72	**< 0,05**	Wind[a]Englisch	0,66	0,30	1,01
			< 0,05[b]	Wind[a]Wiki	0,43	0,02	0,85
			> 0,05	Englisch–Wiki	-0,23	-0,66	0,20
F05	Ist die Suchmaschine präzise?	72	**< 0,05**	Wind[a]Englisch	0,75	0,31	1,19
			> 0,05	Wind–Wiki	0,43	-0,05	0,91
			> 0,05[c]	Englisch–Wiki	-0,32	-0,78	0,14
F06[e]	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	72	**< 0,05**	Wind[a]Englisch	0,68	0,23	1,13
			< 0,05[b]	Wind[a]Wiki	0,52	0,03	1,01
			> 0,05	Englisch–Wiki	-0,16	-0,60	0,28
F09	Ist die Suchmaschine einfach zu bedienen?	72	< 0,05[b]	Wind[a]Englisch	0,23	0,001	0,45
			< 0,05	Wind[a]Wiki	0,32	0,09	0,55
			> 0,05	Englisch–Wiki	0,09	-0,11	0,30
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	72	**< 0,05**	Wind[a]Englisch	1,14	0,66	1,61
			> 0,05	Wind–Wiki	0,39	-0,11	0,89
			< 0,05[b]	Englisch–Wiki[a]	-0,75	-1,35	-0,15
F18[e]	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	72	**< 0,05**	Wind[a]Englisch	0,75	0,31	1,19
			> 0,05	Wind–Wiki	0,30	-0,14	0,73
			> 0,05[c]	Englisch–Wiki	-0,45	-1,01	0,10
F22[e]	Ich bin mit den Suchergebnissen zufrieden.	72	**< 0,05**	Wind[a]Englisch	0,68	0,29	1,08
			> 0,05	Wind–Wiki	0,27	-0,19	0,73
			> 0,05	Englisch–Wiki	-0,41	-0,99	0,17
F25[e]	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	72	**< 0,05**	Wind[a]Englisch	0,55	0,17	0,92
			> 0,05	Wind–Wiki	0,05	-0,25	0,34
			< 0,05[b]	Englisch–Wiki[a]	-0,50	-0,90	-0,10
SK04-M[e]	Suche	72	**< 0,05**	Wind[a]Englisch	0,74	0,40	1,08
			> 0,05	Wind–Wiki	0,41	-0,04	0,86
			> 0,05[c]	Englisch–Wiki	-0,33	-0,84	0,18
SK08-M[e]	Suche	72	**< 0,05**	Wind[a]Englisch	0,58	0,27	0,89
			> 0,05	Wind–Wiki	0,16	-0,19	0,51
			< 0,05[b]	Englisch–Wiki[a]	-0,42	-0,78	-0,07
SK16-M[e]	Aufgabe	72	**< 0,05**	Wind[a]Englisch	0,76	0,29	1,24
			> 0,05	Wind–Wiki	0,24	-0,15	0,63
			> 0,05	Englisch–Wiki	-0,52	-1,05	0,003
SK-A[e]	Accuracy (EUCS)	72	**< 0,05**	Wind[a]Englisch	0,81	0,45	1,16
			< 0,05[b]	Wind[a]Wiki	0,61	0,16	1,06
			> 0,05	Englisch–Wiki	-0,19	-0,68	0,30
SK-C	Content (EUCS)	72	**< 0,05**	Wind[a]Englisch	0,59	0,21	0,96
			> 0,05	Wind–Wiki	0,31	-0,13	0,75
			> 0,05	Englisch–Wiki	-0,28	-0,71	0,15

[a] Bei diesem Thema sind die Probanden zufriedener.

Fortsetzung auf nächster Seite

[b] Nicht mindestens in der Tendenz signifikant.

[c] Eigentlich in der Tendenz signifikant.

[d] Teilweise auch signifikant.

[e] In SP_A mindestens in der Tendenz signifikant.

[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.53 (Fortsetzung)

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
			< 0,05	Wind[a]Englisch	0,52	0,23	0,80
SK-E-88	EUCS-Skala-1988	72	> 0,05	Wind–Wiki	0,25	-0,05	0,54
			> 0,05[c]	Englisch–Wiki	-0,27	-0,59	0,05
			< 0,05	Wind[a]Englisch	0,52	0,17	0,86
SK-E-09[e]	EUCS-Skala-2009	72	> 0,05	Wind–Wiki	0,17	-0,18	0,52
			> 0,05	Englisch–Wiki	-0,35	-0,70	0,01
			< 0,05	Wind[a]Englisch	0,57	0,28	0,86
SK-Z-13[e]	Zusatzskala-2013	72	> 0,05	Wind–Wiki	0,28	-0,09	0,66
			> 0,05	Englisch–Wiki	-0,28	-0,69	0,13
	Wie wahrscheinlich ist es, dass Sie von der		**< 0,05**	Wind[a]Englisch	0,57	0,19	0,95
E04	Leistung, die Sie mithilfe dieser Suchmaschine	72	> 0,05	Wind–Wiki	0,16	-0,20	0,52
	erbringen, sehr überzeugt sind?		> 0,05	Englisch–Wiki	-0,41	-0,87	0,05
			< 0,05	Wind[a]Englisch	0,43	0,11	0,75
E06-M	Erwartungsskala	72	> 0,05	Wind–Wiki	0,16	-0,13	0,46
			> 0,05	Englisch–Wiki	-0,27	-0,63	0,09

[a] Bei diesem Thema sind die Probanden zufriedener.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.54.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung in SP_A unter Ausschluss von SP_{MV}. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
			> 0,05	Wind–Englisch	-1,18	-3,05	0,69
M14[e]	Anz. richtig bew. Dok.	72	**< 0,05**	Wind–Wiki[a]	-2,70	-4,28	-1,13
			> 0,05	Englisch–Wiki	-1,52	-3,22	0,18
			< 0,05[b]	Wind[a]Englisch	0,77	0,14	1,40
M21	Anz. aufg. eher rel. Dok.	72	> 0,05	Wind–Wiki	-0,27	-1,04	0,49
			< 0,05	Englisch–Wiki[a]	-1,05	-1,64	-0,45
			< 0,05	Wind[a]Englisch	-0,55	-1,01	-0,08
M33	Anz. falsch eher rel. bew. rel. Dok.	72	> 0,05	Wind–Wiki	0	-0,43	0,43
			< 0,05[b]	Englisch–Wiki[a]	0,55	0,07	1,02
			< 0,05	Wind–Englisch[a]	0,57	0,05	1,08
M35[e]	Anz. falsch rel. bew. Dok. (4-st.)	72	> 0,05	Wind–Wiki	-0,27	-0,79	0,24
			< 0,05	Englisch[a]Wiki	-0,84	-1,14	-0,54
			> 0,05	Wind–Englisch	0,36	-0,08	0,80
M38	Anz. rel. bew. Dok. (erste 10 Dok.) (4-st.)	72	> 0,05	Wind–Wiki	-0,32	-0,82	0,18
			< 0,05[b]	Englisch–Wiki[a]	-0,68	-1,09	-0,28
			< 0,05	Wind[a]Englisch	1,23	0,44	2,03
B10[e]	Durchschn. Bew. eher rel. Dok.	48	> 0,05	Wind–Wiki	0,33	-0,40	1,07

[a] Dieses Thema entspricht der leichteren Aufgabe.
Fortsetzung auf nächster Seite
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte

Tab. E.54 (Fortsetzung)

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
			$< 0{,}05$[b]	Englisch–Wiki[a]	-0,90	-1,71	-0,09
Z01-log	Durchschn. Betrachtungsz. aller Dok.	72	**$< 0{,}05$**	Wind–Englisch[a]	0,25	0,11	0,39
			$> 0{,}05$[c]	Wind–Wiki	0,06	-0,10	0,23
			$< 0{,}05$[b]	Englisch[a]Wiki	-0,19	-0,37	-0,01
V34[e]	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$	72	$< 0{,}05$[b]	Wind[a]Englisch	0,02	0,003	0,04
			$> 0{,}05$	Wind–Wiki	-0,01	-0,02	0,01
			$< 0{,}05$	Englisch–Wiki[a]	-0,03	-0,04	-0,01
V35[e]	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	72	**$< 0{,}05$**	Wind[a]Englisch	0,04	0,01	0,06
			$> 0{,}05$	Wind–Wiki	-0,005	-0,03	0,02
			$< 0{,}05$	Englisch–Wiki[a]	-0,04	-0,07	-0,01
V37[e]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	**$< 0{,}05$**	Wind–Englisch[a]	-0,15	-0,23	-0,06
			$> 0{,}05$	Wind–Wiki	-0,04	-0,12	0,04
			$< 0{,}05$	Englisch[a]Wiki	0,10	0,02	0,19
V38[e]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$ (4-st.)	72	**$< 0{,}05$**	Wind–Englisch[a]	-0,02	-0,04	-0,005
			$> 0{,}05$	Wind–Wiki	-0,01	-0,03	0,01
			$> 0{,}05$	Englisch–Wiki	0,01	-0,003	0,03
V39	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$ (4-st.)	72	**$< 0{,}05$**	Wind–Englisch[a]	-0,04	-0,06	-0,02
			$> 0{,}05$	Wind–Wiki	-0,02	-0,05	0,01
			$> 0{,}05$[c]	Englisch–Wiki	0,02	-0,01	0,05
V51[e]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	72	**$< 0{,}05$**	Wind–Englisch[a]	0,09	0,03	0,14
			$> 0{,}05$	Wind–Wiki	-0,0001	-0,07	0,07
			$< 0{,}05$	Englisch[a]Wiki	-0,09	-0,15	-0,02
S01[e]	Anz. Suchen	72	**$< 0{,}05$**	Wind–Englisch	-0,75	-1,28	-0,22
			$> 0{,}05$	Wind–Wiki	-0,45	-0,98	0,07
			$> 0{,}05$	Englisch–Wiki	0,30	-0,34	0,94

[a] Dieses Thema entspricht der leichteren Aufgabe.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.55.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit in SP_A unter Ausschluss von SP_{MV}. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
F05	Ist die Suchmaschine präzise?	72	**$< 0{,}05$**	Wind[a]Englisch	0,59	0,14	1,04
			$> 0{,}05$	Wind–Wiki	0,23	-0,29	0,74
			$> 0{,}05$	Englisch–Wiki	-0,36	-0,87	0,14
F06[e]	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	72	**$< 0{,}05$**	Wind[a]Englisch	0,75	0,26	1,24
			$< 0{,}05$[b]	Wind[a]Wiki	0,57	0,05	1,09
			$> 0{,}05$	Englisch–Wiki	-0,18	-0,63	0,26

[a] Bei diesem Thema sind die Probanden zufriedener.
Fortsetzung auf nächster Seite
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.55 (Fortsetzung)

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	72	**< 0,05**	Wind[a]Englisch	0,91	0,41	1,40
			> 0,05	Wind–Wiki	0,27	-0,25	0,80
			< 0,05[b]	Englisch–Wiki[a]	-0,64	-1,21	-0,06
F18[e]	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	72	**< 0,05**	Wind[a]Englisch	0,70	0,27	1,14
			> 0,05	Wind–Wiki	0,41	-0,002	0,82
			> 0,05[c]	Englisch–Wiki	-0,30	-0,83	0,24
F23	Ich bin mit meiner Suchleistung zufrieden.	72	< 0,05[b]	Wind[a]Englisch	0,48	0,07	0,89
			> 0,05	Wind–Wiki	0,02	-0,26	0,31
			< 0,05[b]	Englisch–Wiki[a]	-0,45	-0,87	-0,04
F25[e]	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	72	< 0,05[b]	Wind[a]Englisch	0,61	0,22	1,01
			> 0,05	Wind–Wiki	0,07	-0,21	0,35
			< 0,05	Englisch–Wiki[a]	-0,55	-0,97	-0,12
SK01-M	Genauigkeit	72	**< 0,05**	Wind[a]Englisch	0,67	0,30	1,03
			< 0,05[b]	Wind[a]Wiki	0,43	0,03	0,83
			> 0,05	Englisch–Wiki	-0,23	-0,67	0,20
SK04-M[e]	Suche	72	< 0,05[b]	Wind[a]Englisch	0,83	0,45	1,20
			> 0,05	Wind–Wiki	0,22	-0,18	0,62
			< 0,05[b]	Englisch–Wiki[a]	-0,61	-1,10	-0,11
SK08-M[e]	Suche	72	**< 0,05**	Wind[a]Englisch	0,46	0,12	0,80
			> 0,05	Wind–Wiki	0,05	-0,37	0,46
			< 0,05[b]	Englisch–Wiki[a]	-0,41	-0,83	-0,0002
SK16-M[e]	Aufgabe	72	**< 0,05**	Wind[a]Englisch	0,73	0,29	1,17
			> 0,05	Wind–Wiki	0,25	-0,24	0,74
			> 0,05[c]	Englisch–Wiki	-0,48	-1,01	0,05
SK-A[e]	Accuracy (EUCS)	72	**< 0,05**	Wind[a]Englisch	0,67	0,26	1,08
			> 0,05[c]	Wind–Wiki	0,20	-0,23	0,64
			< 0,05[b]	Englisch–Wiki[a]	-0,47	-0,88	-0,05
SK-E-13[e]	EUCS-Skala-2013	72	**< 0,05**	Wind[a]Englisch	0,53	0,19	0,87
			> 0,05	Wind–Wiki	0,24	-0,15	0,64
			> 0,05	Englisch–Wiki	-0,28	-0,64	0,08
E06-M	Erwartung	72	< 0,05[b]	Wind[a]Englisch	0,35	0,04	0,65
			> 0,05	Wind–Wiki	0,09	-0,16	0,34
			> 0,05	Englisch–Wiki	-0,26	-0,59	0,08

[a] Bei diesem Thema sind die Probanden zufriedener.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.56.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung in SP_A unter Ausschluss von SP_{SB}. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
M14[e]	Anz. richtig bew. Dok.	48	$> 0{,}05$	Wind–Englisch	-0,87	-3,05	1,32
			$\mathbf{< 0{,}05}$	Wind–Wiki[a]	-2,87	-4,82	-0,91
			$> 0{,}05$	Englisch–Wiki	-2	-4,28	0,28
M21	Anz. aufg. eher rel. Dok.	48	$< 0{,}05$[b]	Wind[a]Englisch	1,23	0,40	2,06
			$> 0{,}05$	Wind–Wiki	0,10	-0,79	0,99
			$\mathbf{< 0{,}05}$	Englisch–Wiki[a]	-1,13	-2,01	-0,25
M35[e]	Anz. falsch rel. bew. Dok. (4-st.)	48	$< 0{,}05$[b]	Wind–Englisch[a]	0,57	0,18	0,96
			$> 0{,}05$	Wind–Wiki	-0,27	-0,84	0,31
			$\mathbf{< 0{,}05}$	Englisch[a]Wiki	-0,83	-1,40	-0,27
M45	Anz. richtig rel. bew. Dok. (4-st.)	48	$\mathbf{< 0{,}05}$	Wind–Englisch[a]	-0,80	-1,29	-0,31
			$> 0{,}05$	Wind–Wiki	-0,30	-0,97	0,37
			$> 0{,}05$	Englisch–Wiki	0,50	-0,24	1,24
V13	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$	48	$> 0{,}05$	Wind–Englisch	-0,19	-0,38	0,003
			$\mathbf{< 0{,}05}$	Wind–Wiki[a]	-0,25	-0,46	-0,04
			$> 0{,}05$[d]	Englisch–Wiki	-0,06	-0,27	0,15
V35[e]	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	48	$> 0{,}05$	Wind–Englisch	0,03	-0,002	0,06
			$> 0{,}05$	Wind–Wiki	-0,01	-0,05	0,02
			$\mathbf{< 0{,}05}$	Englisch–Wiki[a]	-0,04	-0,07	-0,01
V37[e]	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	24	$< 0{,}05$[b]	Wind–Englisch[a]	-0,17	-0,31	-0,03
			$> 0{,}05$	Wind–Wiki	-0,06	-0,16	0,05
			$> 0{,}05$[d]	Englisch–Wiki	0,11	-0,04	0,26
V47	$\frac{\text{Anz. falsch eher rel. bew. rel. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	24	$\mathbf{< 0{,}05}$	Wind[a]Englisch	-0,54	-0,82	-0,27
			$> 0{,}05$	Wind–Wiki	-0,15	-0,35	0,05
			$< 0{,}05$[b]	Englisch–Wiki[a]	0,39	0,08	0,69
V51[e]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	48	$\mathbf{< 0{,}05}$	Wind–Englisch[a]	0,12	0,06	0,18
			$> 0{,}05$	Wind–Wiki	0,0001	-0,08	0,08
			$\mathbf{< 0{,}05}$	Englisch[a]Wiki	-0,12	-0,19	-0,04

[a] Dieses Thema entspricht der leichteren Aufgabe.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.57.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerleistung in SP_A unter Ausschluss von SP_{IZ}. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
V51[e]	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	48	$\mathbf{< 0{,}05}$	Wind-Englisch[a]	0,08	0,02	0,15
			$> 0{,}05$	Wind-Wiki	0,0005	-0,07	0,07
			$< 0{,}05$[b]	Englisch[a]Wiki	-0,08	-0,16	-0,01

[a] Dieses Thema entspricht der leichteren Aufgabe.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

Tab. E.58.: Signifikante Ergebnisse der robusten Varianzanalyse zur Untersuchung des Einflusses von Topiceffekten auf die Benutzerzufriedenheit in SP_A unter Ausschluss von SP_{IZ}. Es werden nur mindestens in der Tendenz signifikante Topiceffekte berichtet, wobei jeweils das signifikanteste Ergebnis ausgewählt wird. Fett hervorgehoben sind Effekte, die in mindestens vier von fünf Stichproben nachweisbar sind. n bezeichnet die Anzahl an Fällen pro Topic.

ID	Beschreibung	n	p	Vergleich	d	lwr	upr
F18[e]	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	48	$\mathbf{< 0{,}05}$	Wind[a]Englisch	0,83	0,39	1,28
			$> 0{,}05$	Wind–Wiki	0,10	-0,13	0,33
			$< 0{,}05$[b]	Englisch–Wiki[a]	-0,73	-1,15	-0,32
SK16-M[e]	Aufgabe	48	$< 0{,}05$[b]	Wind[a]Englisch	0,87	0,40	1,34
			$> 0{,}05$	Wind–Wiki	0,10	-0,29	0,49
			$< 0{,}05$[b]	Englisch–Wiki[a]	-0,77	-1,32	-0,21

[a] Bei diesem Thema sind die Probanden zufriedener.
[b] Nicht mindestens in der Tendenz signifikant.
[c] Eigentlich in der Tendenz signifikant.
[d] Teilweise auch signifikant.
[e] In SP_A mindestens in der Tendenz signifikant.
[f] Robuste Varianzanalyse konnte nicht für alle Stichproben signifikante Topiceffekte nachweisen.

E.6. Weitere Ergebnisse der Kovarianzanalyse

In diesem Anhang sind weitere Befunde der im dritten Experiment durchgeführten Kovarianzanalysen aufgeführt. Um die Tabellen zu den einzelnen Aspekten leichter auffindbar zu machen, ist auch dieser Anhang in drei Unterabschnitte untergliedert: In Abschnitt E.6.1 wird analog zum zweiten Experiment zunächst eine Übersicht über Variablen gegeben, für die die im Kontext der Hauptauswertung nachgewiesenen Befunde auch unter Einbeziehung der Kovariaten stabil bleiben. Abschnitt E.6.2 hingegen stellt die Gruppenmittelwerte zu den im Hauptteil der Arbeit beschriebenen Ergebnissen in Bezug auf das Hinzukommen oder Verschwinden von Haupteffekten oder Interaktionen bereit. In Abschnitt E.6.3 schließlich sind die entsprechenden Teststatistiken wiedergegeben.

E.6.1. Variablen mit stabilen Befunden in Bezug auf demographische und erfahrungsbezogene Kovariaten

Dieser Abschnitt stellt eine Übersicht derjenigen Variablen bereit, die im Rahmen der Kovarianzanalyse des dritten Experiments über alle untersuchten demographischen und erfahrungsbezogenen Einflussgrößen hinweg ein mit den Ergebnissen der Hauptauswertung übereinstimmendes Befundmuster aufweisen. Aufgrund der hohen Anzahl der Variablen sind die Ergebnisse getrennt nach Benutzerleistung und Benutzerzufriedenheit zusammengefasst.

Tab. E.59.: Übersicht über Benutzerleistungsvariablen, für die im Rahmen der Kovarianzanalyse in SP_A keine Effekte neu hinzukommen oder verschwinden

ID	Beschreibung	n_{min}	n_{max}
M01	Anz. aufg. Dok.	29	29
M03	Anz. aufg. Dok. (erste Suche)	29	29
M04	Anz. aufg. Dok. (letzte Suche)	29	29
M05	Anz. aufg. irrel. Dok.	29	29
M06	Anz. aufg. rel. Dok.	29	29
M07	Anz. falsch irrel. bew. Dok.	29	29
M08	Anz. falsch rel. bew. Dok.	29	29
M09	Anz. irrel. bew. Dok.	29	29
M12	Anz. rel. bew. Dok. (erste Suche)	29	29
M13	Anz. rel. bew. Dok. (letzte Suche)	29	29
M14	Anz. richtig bew. Dok.	29	29
M16	Anz. richtig rel. bew. Dok.	29	29
M20	Anz. aufg. eher irrel. Dok.	29	29
M24	Anz. eher irrel. bew. Dok.	29	29
M25	Anz. eher rel. bew. Dok.	29	29
M30	Anz. falsch eher rel. bew. Dok.	29	29
M34	Anz. falsch irrel. bew. Dok. (4-st.)	29	29
M35	Anz. falsch rel. bew. Dok. (4-st.)	29	29
M28	Anz. falsch eher irrel. bew. irrel. Dok.	29	29
M32	Anz. falsch eher rel. bew. irrel. Dok.	29	29
M37	Anz. rel. bew. Dok. (4-st.)	29	29
M39	Anz. rel. bew. Dok. (erste Suche) (4-st.)	29	29
M40	Anz. rel. bew. Dok. (letzte Suche) (4-st.)	29	29
M33	Anz. falsch eher rel. bew. rel. Dok.	29	29
M41	Anz. richtig bew. Dok. (4-st.)	29	29
M43	Anz. richtig eher rel. bew. Dok.	29	29
M47	Anz. richtig rel. bew. Dok. (erste Suche) (4-st.)	29	29
M46	Anz. richtig rel. bew. Dok. (erste 10 Dok.) (4-st.)	29	29
M48	Anz. richtig rel. bew. Dok. (letzte Suche) (4-st.)	29	29
B05	Durchschn. Bew. rel. Dok. (erste Suche)	12	12
B04	Durchschn. Bew. rel. Dok.	29	29
B10	Durchschn. Bew. eher rel. Dok.	13	13
B16	Durchschn. Bew. rel. Dok. (4-st.)	29	29
Z01	Durchschn. Betrachtungsz. aller Dok.	29	29
Z02-log	Durchschn. Betrachtungsz. falsch bew. Dok.	20	20
Z03	Durchschn. Betrachtungsz. falsch irrel. bew. Dok.	10	10
Z05	Durchschn. Betrachtungsz. irrel. bew. Dok.	22	22
Z07-log	Durchschn. Betrachtungsz. rel. bew. Dok.	27	27
Z08	Durchschn. Betrachtungsz. rel. Dok.	29	29
Z08-log	Durchschn. Betrachtungsz. rel. Dok.	29	29
Z09	Durchschn. Betrachtungsz. richtig bew. Dok.	28	28
Z11-log	Durchschn. Betrachtungsz. richtig rel. bew. Dok.	27	27
Z22-log	Durchschn. Betrachtungsz. rel. bew. Dok. (4-st.)	20	20
Z23	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	29	29

Fortsetzung auf nächster Seite

Tab. E.59 (Fortsetzung)

ID	Beschreibung	n_{min}	n_{max}
Z23-log	Durchschn. Betrachtungsz. rel. Dok. (4-st.)	29	29
Z24	Durchschn. Betrachtungsz. richtig bew. Dok. (4-st.)	22	22
Z24-log	Durchschn. Betrachtungsz. richtig bew. Dok. (4-st.)	22	22
V31/BP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	27	27
V32/BR	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	27	27
V01	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$	29	29
V02	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$	29	29
V03	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$	29	29
V04	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$	29	29
V05	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	29	29
V06	$\frac{\text{Anz. falsch irrel. bew. Dok.}}{\text{Anz. irrel. bew. Dok.}}$	22	22
V08	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	29	29
V09	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$	27	27
V10	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. richtig rel. bew. Dok.}}$	27	27
V11	$\frac{\text{Anz. irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	29	29
V12	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	29	29
V13	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$	29	29
V24	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. aufg. Dok. (letzte Suche)}}$	20	20
V26	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. rel. Dok. im Korpus}}$	20	20
V27	$\frac{\text{Anz. richtig rel. bew. Dok. (letzte Suche)}}{\text{Anz. zurückgeg. rel. Dok. (letzte Suche)}}$	20	20
V28/PCP	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	27	27
V35	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	29	29
V36	$\frac{\text{Anz. aufg. irrel. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	29	29
V37	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	29	29
V38	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. rel. Dok. im Korpus}}$ (4-st.)	29	29
V39	$\frac{\text{Anz. aufg. rel. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$ (4-st.)	29	29
V45	$\frac{\text{Anz. falsch eher rel. bew. eher irrel. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	17	17
V40	$\frac{\text{Anz. falsch eher irrel. bew. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	13	13
V44	$\frac{\text{Anz. falsch eher rel. bew. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	17	17
V51	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	29	29
V52	$\frac{\text{Anz. falsch rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$ (4-st.)	20	20
V42	$\frac{\text{Anz. falsch eher irrel. bew. irrel. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	13	13
V55	$\frac{\text{Anz. rel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	29	29
V43	$\frac{\text{Anz. falsch eher irrel. bew. rel. Dok.}}{\text{Anz. eher irrel. bew. Dok.}}$	13	13
V47	$\frac{\text{Anz. falsch eher rel. bew. rel. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	17	17
V56	$\frac{\text{Anz. richtig bew. Dok.}}{\text{Anz. aufg. Dok.}}$ (4-st.)	29	29

Fortsetzung auf nächster Seite

Tab. E.59 (Fortsetzung)

ID	Beschreibung	n_{min}	n_{max}
V59	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$	29	29
V60	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. zurückgeg. eher rel. Dok.}}$	29	29
V78	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. rel. bew. Dok.}}$ (4-st.)	14	14
V80	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. zurückgeg. rel. Dok.}}$ (4-st.)	14	14
S01	Anz. Suchen	29	29
S02	Erste betr. Rankingpos.	29	29
S03	Letzte betr. Rankingpos.	29	29

Tab. E.60.: Übersicht über Benutzerzufriedenheitsvariablen, für die im Rahmen der Kovarianzanalyse in SP_A keine Effekte neu hinzukommen oder verschwinden.

ID	Beschreibung	n_{min}	n_{max}
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	29	29
F04	Liefert die Suchmaschine genügend Information?	29	29
F09	Ist die Suchmaschine einfach zu bedienen?	29	29
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	29	29
F13	Sind Sie mit der Suchmaschine zufrieden?	29	29
F14	Es war einfach, die Aufgabe zu bearbeiten.	29	29
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	29	29
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	29	29
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	29	29
F22	Ich bin mit den Suchergebnissen zufrieden.	29	29
F23	Ich bin mit meiner Suchleistung zufrieden.	29	29
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	29	29
SK08-M	Suche	29	29
SK11-M[a]	Eigenleistung	29	29
SK13-F	Benutzerfreundlichkeit	29	29
SK16-F	Aufgabe	29	29
SK16-M	Aufgabe	29	29
SK19-F	Eigenleistung	29	29
SK-A	Accuracy (EUCS)	29	29
SK-C	Content (EUCS)	29	29
SK-E[b]	Ease of Use (EUCS)	29	29
SK-E-88	EUCS-Skala-1988	29	29
SK-E-13	EUCS-Skala-2013	29	29
SK-G-13	Gesamtskala-2013	29	29
SK-T	Timeliness (EUCS)	29	29
SK-Z-13	Zusatzskala-2013	29	29
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	29	29
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	29	29
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	29	29

[a] Entspricht auch den Skalen SK15-M und SK19-M.

[b] Entspricht auch der Skala SK13-M.

E.6.2. Gruppenmittelwerte der Kovarianzanalyse

Dieser Abschnitt beinhaltet die innerhalb der Arbeit aus Platzgründen nicht berichteten Mittelwerte der signifikanten Kovarianzanalysen für die eine Änderung der im Kontext der Hauptanalyse beschriebenen Effekte zu beobachten ist. Analog zu den in Abschnitt 7.4.6.2 dargestellten Übersichtstabellen sind die ermittelten Gruppenmittelwerte in drei verschiedenen Tabellen zusammengestellt. Während Tabelle E.61 die entsprechenden Werte der Benutzerleistunganalyse ausweist, sind die Ergebnisse bezüglich der Benutzerzufriedenheit getrennt nach den beiden im Rahmen der Zufriedenheitsanalyse betrachteten Kovariatengruppen aufgeführt.

Tab. E.61.: Signifikante Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses demographischer und erfahrungsbezogener Störfaktoren auf die Benutzerleistung in SP_A.

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
M02	Anz. aufg. Dok. (erste 10 Dok.)	K02	2,85	2,52	2,85	2,52	**3,33**[a]	**2,37**	**2,37**	**2,67**
		K03	2,67	2,33	2,50	2,50	**3,00**[a]	**2,33**	**2,00**	**2,67**
		K04	3,00	2,50	2,83	2,67	**3,33**[a]	**2,67**	**2,33**	**2,67**
		K08	2,83	2,50	2,83	2,50	**3,33**[a]	**2,33**	**2,33**	**2,67**
		K09	3,00	2,54	2,83	2,71	**3,33**[a]	**2,67**	**2,33**	**2,74**
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	K03	1,33	1,17	1,33	1,17	**1,67**[b]	**1,00**	**1,00**	**1,33**
		K09	1,25	1,02	1,27	1,00	**1,59**[b]	**0,92**	**0,96**	**1,09**
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	K02	1,67	1,42	1,67	1,42	**2,00**[b]	**1,33**	**1,33**	**1,50**
		K08	**1,60**[b]	**1,26**	1,59	1,26	1,93	1,27	1,26	1,26
M21	Anz. aufg. eher rel. Dok.	K09	**2,29**[b]	**1,69**	2,00	1,98	2,33	2,25	1,66	1,71
		K10	**2,36**[b]	**1,74**	2,07	2,02	2,40	2,31	1,74	1,73
M26	Anz. falsch eher irrel. bew. Dok.	K02	**1,42**[a]	**1,08**	1,25	1,25	1,50	1,33	1,00	1,17
		K05	**1,11**[a]	**0,87**	0,94	1,03	1,21	1,00	0,67	1,06
		K08	**1,58**[a]	**1,33**	1,42	1,50	1,67	1,50	1,17	1,50
M29	Anz. falsch eher irrel. bew. rel. Dok.	K03	0,67	0,50	**0,50**[b]	**0,67**	0,67	0,67	0,33	0,67
		K04	0,67	0,50	**0,50**[b]	**0,67**	0,67	0,67	0,33	0,67
		K09	0,67	0,50	**0,50**[b]	**0,67**	0,67	0,67	0,33	0,67
		K10	0,59	0,51	**0,42**[b]	**0,67**	0,48	0,69	0,37	0,65
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	K10	**5,55**[a]	**5,87**	5,78	5,64	5,50	5,60	6,06	5,69
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	K05	**41,61**[a]	**36,67**	39,89	38,38	43,69	39,52	36,10	37,24
		K08	**44,05**[a]	**39,12**	43,11	40,06	47,33	40,77	38,89	39,35
		K10	**41,40**[a]	**36,44**	39,77	38,07	44,03	38,77	35,51	37,36
Z05-log	Durchschn. Betrachtungsz. irrel. bew. Dok.	K02	**3,40**[a]	**3,24**	3,30	3,33	3,43	3,36	3,18	3,30
		K10	**3,46**[a]	**3,30**	3,35	3,40	3,49	3,43	3,22	3,38
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	K04	0,14	0,23	**0,17**[a]	**0,20**	0,12	0,16	0,22	0,24

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

Fortsetzung auf nächster Seite

Tab. E.61 (Fortsetzung)

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
V29	Anz. richtig rel. bew. Dok. / Anz. aufg. rel. Dok.	K02	**0,71**[b]	**0,77**	0,74	0,74	0,70	0,71	0,78	0,77
		K03	**0,71**[b]	**0,78**	0,75	0,75	0,71	0,72	0,79	0,78
		K04	**0,66**[b]	**0,73**	0,70	0,69	0,66	0,66	0,74	0,72
		K05	**0,67**[b]	**0,75**	0,71	0,70	0,68	0,65	0,75	0,75
		K08	**0,67**[b]	**0,74**	0,70	0,70	0,66	0,67	0,74	0,73
		K09	**0,64**[b]	**0,71**	0,68	0,68	0,64	0,64	0,72	0,71
		K10	**0,67**[b]	**0,74**	0,71	0,70	0,67	0,67	0,75	0,72
V34	Anz. aufg. eher rel. Dok. / Anz. eher rel. Dok. im Korpus	K05	**0,06**[b]	**0,05**	0,06	0,06	0,06	0,06	0,05	0,05
		K09	**0,06**[b]	**0,05**	0,05	0,05	0,06	0,06	0,04	0,05
		K10	**0,06**[b]	**0,05**	0,05	0,06	0,06	0,06	0,05	0,05
V58	Anz. richtig eher rel. bew. Dok. / Anz. eher rel. bew. Dok.	K02	0,33	0,29	**0,27**[a]	**0,35**	0,27	0,39	0,27	0,31

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

Tab. E.62.: Signifikante Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses demographischer und erfahrungsbezogener Störfaktoren auf die Benutzerzufriedenheit in SP_A.

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F11	Liefert die Suchmaschine aktuelle Information?	K03	3,62	3,51	**3,68**[b]	**3,45**	3,68	3,55	3,67	3,34
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	K02	**3,08**[a]	**2,83**	3,25	2,67	3,33	2,83	3,17	2,50
SK03-M	Benutzerfreundlichkeit	K05	3,60	3,41	3,76	3,25	**3,70**[b]	**3,50**	**3,82**	**3,00**
SK14-M	Suche	K02	**3,44**[b]	**3,17**	3,50	3,11	3,61	3,28	3,39	2,94
		K05	**3,39**[b]	**3,11**	3,50	3,00	3,56	3,22	3,44	2,78
SK15-F	Eigenleistung	K02	−0,05	−0,08	**0,20**[b]	**−0,34**	0,19	−0,29	0,22	−0,38
		K03	0,21	0,19	**0,47**[b]	**−0,08**	0,43	−0,02	0,51	−0,14
		K05	0,05	$5{\cdot}10^{-4}$	**0,29**[b]	**−0,25**	0,28	−0,19	0,31	−0,30
		K09	−0,26	−0,29	**−0,01**[b]	**−0,54**	−0,02	−0,49	$-6{\cdot}10^{-4}$	−0,59
		K10	−0,03	−0,06	**0,22**[b]	**−0,32**	0,21	−0,27	0,24	−0,37
E01	Ich glaube, ich werde in zehn Minuten ... relevante Dokumente finden.	K04	4,50	5,00	**5,17**[b]	**4,33**	5,00	4,00	5,33	4,67
		K10	4,14	4,58	**4,86**[b]	**3,86**	4,72	3,56	5,00	4,17

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

Tab. E.63.: **Fertig** Neu Signifikante Ergebnisse der Kovarianzanalyse zur Untersuchung des Einflusses leistungsbezogener Störfaktoren auf die Benutzerzufriedenheit in SP_A.

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	B11	2,75	2,61	**2,85**[a]	**2,52**	2,79	2,71	2,90	2,33
		B17	**1,15**[b]	**0,77**	1,23	0,69	1,30	1,00	1,17	0,38
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	B01	3,18	2,87	**3,12**[a]	**2,93**	3,34	3,02	2,90	2,84
		B04	**−0,002**[b]	**−0,31**	0,01	−0,32	0,17	−0,17	−0,14	−0,48
		B05	**1,58**[b]	**1,17**	1,57	1,18	1,72	1,44	1,42	0,92
		B16	**0,09**[b]	**−0,21**	0,10	−0,22	0,24	−0,06	−0,03	−0,39
		B17	**1,51**[b]	**1,17**	1,55	1,12	1,65	1,36	1,46	0,88
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	B04	**0,24**[b]	**−0,11**	0,24	−0,12	0,33	0,14	0,16	−0,39
		B10	1,66	1,48	**1,66**[a]	**1,48**	1,76	1,56	1,57	1,40
		B11	1,65	1,65	**1,80**[a]	**1,50**	1,71	1,60	1,89	1,40
		B16	**0,48**[b]	**0,15**	0,52	0,12	0,60	0,35	0,43	−0,12
F04	Liefert die Suchmaschine genügend Information?	B10	2,49	2,36	**2,60**[a]	**2,25**	2,68	2,30	2,52	2,20
		B11	1,63	1,26	**1,50**[a]	**1,39**	1,66	1,60	1,33	1,18
F05	Ist die Suchmaschine präzise?	B04	**0,47**[b]	**0,05**	0,61	−0,08	0,77	0,18	0,44	−0,34
		B06	**0,76**[b]	**0,30**	0,82	0,25	0,97	0,56	0,67	−0,06
		B12	**2,41**[b]	**1,87**	2,35	1,92	2,45	2,36	2,26	1,47
		B16	**0,40**[b]	**−0,07**	0,46	−0,14	0,69	0,11	0,24	−0,39
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	B12	**1,90**[b]	**1,39**	1,88	1,40	1,98	1,81	1,78	1,00
		B16	**−0,20**[b]	**−0,59**	−0,07	−0,72	0,03	−0,44	−0,18	−1,00
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	B05	1,75	1,58	**1,79**[a]	**1,53**	1,76	1,73	1,83	1,33
		B10	2,23	2,11	**2,32**[a]	**2,02**	2,28	2,18	2,37	1,85
		B11	2,59	2,14	**2,47**[a]	**2,26**	2,54	2,65	2,40	1,87
F08	Ist die Suchmaschine benutzerfreundlich?	S06	4,17	3,83	4,33	3,67	**4,33**[b]	**4,00**	**4,33**	**3,33**
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	B10	3,16	3,18	**3,29**[a]	**3,04**	3,33	2,98	3,25	3,10
		B11	4,33	4,33	**4,33**[a]	**4,33**	4,33	4,33	4,33	4,33
		B12	2,82	2,53	**2,80**[a]	**2,55**	2,91	2,72	2,69	2,38
F12	Ist die Suchmaschine erfolgreich?	M10	2,98	2,93	3,28	2,63	**3,13**[b]	**2,83**	**3,43**	**2,43**
		M37	3,12	3,02	3,40	2,74	**3,29**[b]	**2,95**	**3,52**	**2,52**
F13	Sind Sie mit der Suchmaschine zufrieden?	B06	**0,14**[b]	**−0,22**	0,21	−0,29	0,30	−0,01	0,12	−0,56
		B11	3,32	3,14	**3,50**[a]	**2,95**	3,53	3,10	3,47	2,80
		B16	**0,19**[b]	**−0,10**	0,32	−0,24	0,39	−0,01	0,26	−0,47
F14	Es war einfach, die Aufgabe zu bearbeiten.	B06	2,07	1,84	**2,03**[a]	**1,88**	2,21	1,92	1,85	1,83
		B11	2,49	2,33	**2,43**[a]	**2,39**	2,71	2,27	2,15	2,52
		B12	2,71	2,38	**2,65**[a]	**2,45**	2,93	2,49	2,36	2,40
		B17	1,88	1,71	**1,89**[a]	**1,70**	2,01	1,76	1,77	1,64
		B18	2,09	1,78	**2,06**[a]	**1,81**	2,30	1,88	1,83	1,74

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Entspricht auch der Skala SK18-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Tab. E.63 (Fortsetzung)

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	B05	**2,26**[b]	**1,93**	2,30	1,89	2,35	2,17	2,25	1,61
		B11	3,37	3,29	**3,35**[a]	**3,32**	3,24	3,51	3,47	3,12
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	B05	2,05	1,62	**1,89**[a]	**1,78**	1,98	2,12	1,81	1,43
		B06	**1,19**[b]	**0,75**	1,15	0,80	1,31	1,08	0,99	0,51
		B10	2,19	1,81	**2,11**[a]	**1,88**	2,34	2,04	1,89	1,72
		B12	**2,63**[b]	**1,94**	2,55	2,02	2,78	2,49	2,32	1,56
		B12	2,48	1,90	**2,39**[a]	**1,99**	2,65	2,31	2,14	1,66
		B16	**0,87**[b]	**0,41**	0,90	0,39	1,06	0,68	0,74	0,09
		B17	1,93	1,56	**1,95**[a]	**1,55**	2,00	1,86	1,89	1,23
		B18	1,31	0,94	**1,33**[a]	**0,92**	1,45	1,18	1,21	0,66
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	B04	**1,34**[b]	**0,96**	1,30	1,00	1,54	1,14	1,06	0,86
		B05	**1,40**[b]	**0,98**	1,35	1,03	1,55	1,25	1,14	0,81
		B06	**1,58**[b]	**1,31**	1,58	1,31	1,74	1,42	1,41	1,20
		B12	**2,06**[b]	**1,67**	2,01	1,73	2,26	1,87	1,76	1,58
		B16	**0,98**[b]	**0,55**	0,92	0,61	1,16	0,79	0,67	0,43
		B18	**1,57**[b]	**1,23**	1,55	1,26	1,79	1,36	1,31	1,16
		B18	1,70	1,48	**1,70**[a]	**1,49**	1,91	1,49	1,49	1,48
		S06	**3,90**[b]	**3,57**	3,92	3,56	4,08	3,72	3,75	3,39
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	B05	**1,18**[a]	**0,91**	1,25	0,84	1,36	1,00	1,13	0,68
		B10	**1,53**[a]	**1,26**	1,62	1,17	1,75	1,31	1,50	1,02
		B12	**1,94**[a]	**1,62**	2,01	1,55	2,08	1,79	1,94	1,31
		B12	2,01	1,55	**1,97**[a]	**1,58**	2,04	1,97	1,90	1,20
		B17	1,53	1,11	**1,49**[a]	**1,15**	1,59	1,47	1,39	0,83
		S03	**3,33**[a]	**3,12**	3,48	2,97	3,52	3,15	3,45	2,79
		S04	**2,65**[a]	**2,43**	2,81	2,27	2,89	2,42	2,73	2,12
		S06	**3,42**[a]	**3,19**	3,59	3,02	3,60	3,24	3,57	2,81
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	B05	0,22	0,10	**0,33**[a]	**−0,01**	0,31	0,13	0,35	−0,16
		B10	1,10	1,16	**1,22**[a]	**1,03**	1,31	0,88	1,14	1,19
		B17	0,10	−0,03	**0,16**[a]	**−0,09**	0,17	0,03	0,16	−0,21
		B18	0,79	0,79	**0,94**[a]	**0,64**	1,00	0,58	0,89	0,69
F22	Ich bin mit den Suchergebnissen zufrieden.	B04	**−0,05**[b]	**−0,37**	0,02	−0,45	0,16	−0,26	−0,11	−0,64
		B05	**1,08**[b]	**0,63**	1,06	0,66	1,19	0,97	0,92	0,34
		B11	**2,05**[b]	**1,18**	1,88	1,35	2,12	1,98	1,63	0,72
		B11	2,98	2,82	**3,12**[a]	**2,68**	3,30	2,67	2,95	2,69
		B12	**2,26**[b]	**1,73**	2,22	1,78	2,43	2,09	2,00	1,46
		B16	**0,31**[b]	**−0,10**	0,32	−0,10	0,46	0,16	0,17	−0,37

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Entspricht auch der Skala SK18-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Tab. E.63 (Fortsetzung)

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
F23	Ich bin mit meiner Suchleistung zufrieden.	B05	1,46	1,37	**1,54**[a]	**1,28**	1,55	1,36	1,54	1,19
		B06	1,36	1,32	**1,44**[a]	**1,25**	1,48	1,25	1,40	1,24
		B10	1,96	1,80	**1,97**[a]	**1,79**	2,02	1,91	1,91	1,68
		B11	3,23	3,00	**3,39**[a]	**2,84**	3,43	3,03	3,36	2,65
		B16	1,28	1,15	**1,34**[a]	**1,09**	1,41	1,16	1,26	1,03
		B17	1,34	1,31	**1,44**[a]	**1,21**	1,38	1,30	1,51	1,11
		B18	1,60	1,57	**1,74**[a]	**1,43**	1,73	1,46	1,76	1,39
		M10	2,90	2,97	**3,04**[a]	**2,83**	3,01	2,79	3,06	2,87
		M45	3,04	2,92	**3,08**[a]	**2,88**	3,17	2,92	3,00	2,83
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	B05	**1,96**[b]	**1,33**	1,93	1,36	2,18	1,74	1,68	0,99
		B05	2,54	2,18	**2,50**[a]	**2,21**	2,64	2,43	2,35	2,00
		B11	0,70	0,28	**0,65**[a]	**0,32**	0,91	0,49	0,40	0,15
		B12	**2,89**[b]	**2,21**	2,84	2,26	3,10	2,69	2,59	1,83
		B16	**1,32**[b]	**0,95**	1,31	0,97	1,47	1,18	1,16	0,75
		B17	**1,91**[b]	**1,37**	1,88	1,40	1,99	1,83	1,77	0,97
		B18	**1,63**[b]	**1,17**	1,64	1,17	1,83	1,43	1,44	0,91
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	B01	2,79	2,36	**2,74**[a]	**2,41**	3,09	2,49	2,39	2,34
		B11	2,30	1,93	**2,25**[a]	**1,98**	2,27	2,32	2,22	1,65
		B16	**0,28**[b]	**−0,05**	0,37	−0,14	0,44	0,12	0,30	−0,40
		B17	1,23	0,83	1,34	0,72	**1,30**[b]	**1,16**	**1,39**	**0,28**
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	B03	2,41	1,47	**2,20**[a]	**1,68**	2,70	2,12	1,70	1,24
		B16	**−0,006**[b]	**−0,43**	0,09	−0,52	0,18	−0,19	−0,007	−0,84
		B17	−0,07	−0,47	**−0,08**[a]	**−0,46**	−0,07	−0,06	−0,08	−0,86
SK01-M	Genauigkeit	B11	2,27	2,10	**2,43**[a]	**1,94**	2,46	2,08	2,40	1,80
		B16	**0,27**[b]	**−0,12**	0,34	−0,19	0,49	0,05	0,19	−0,42
SK02-M	Inhalt	B01	2,77	2,76	**2,96**[a]	**2,57**	3,18	2,35	2,74	2,78
		B05	**1,79**[b]	**1,45**	1,86	1,38	1,98	1,60	1,75	1,16
		B11	2,73	2,68	**2,72**[a]	**2,68**	2,75	2,70	2,69	2,67
SK03-M	Benutzerfreundlichkeit	B04	**2,05**[b]	**1,68**	2,10	1,62	2,16	1,93	2,04	1,32
		B16	**1,81**[b]	**1,49**	1,93	1,37	1,98	1,65	1,89	1,10
		B17	2,54	2,23	2,64	2,13	**2,53**[b]	**2,56**	**2,75**	**1,71**
		B18	**2,09**[b]	**1,75**	2,20	1,63	**2,19**[b]	**1,99**	**2,22**	**1,28**
		M45	**3,48**[b]	**3,18**	3,56	3,11	**3,54**[b]	**3,43**	**3,58**	**2,79**
		S03	3,59	3,36	3,75	3,20	**3,70**[b]	**3,48**	**3,80**	**2,91**

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Entspricht auch der Skala SK18-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Tab. E.63 (Fortsetzung)

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
SK04-M	Suche	B04	**0,76**[b]	**0,35**	0,75	0,35	0,94	0,58	0,56	0,13
		B05	**1,57**[b]	**1,13**	1,55	1,16	1,69	1,45	1,40	0,87
		B06	**1,43**[b]	**1,14**	1,50	1,08	1,63	1,23	1,36	0,92
		B12	**2,27**[b]	**1,78**	2,19	1,86	2,38	2,17	2,01	1,55
		B16	**0,88**[b]	**0,48**	0,87	0,49	1,03	0,73	0,71	0,25
		B17	**1,40**[b]	**0,93**	1,39	0,95	1,48	1,32	1,30	0,57
		B18	**1,45**[b]	**1,14**	1,50	1,09	1,60	1,29	1,39	0,89
		M45	**3,05**[b]	**2,77**	3,11	2,71	3,22	2,88	3,01	2,53
SK07-M[c]	Benutzerfreundlichkeit	B01	3,33	3,25	**3,42**[a]	**3,17**	3,50	3,17	3,33	3,17
		B04	**1,24**[b]	**0,72**	1,20	0,76	1,40	1,07	1,01	0,44
		B06	**1,66**[b]	**1,25**	1,66	1,24	1,79	1,53	1,53	0,96
		B12	**2,57**[b]	**1,83**	2,42	1,97	2,64	2,49	2,20	1,46
		B16	**1,24**[b]	**0,82**	1,26	0,80	1,40	1,08	1,13	0,51
		B17	**2,07**[b]	**1,47**	2,01	1,52	2,16	1,97	1,86	1,07
		B18	**1,52**[b]	**1,07**	1,56	1,03	1,72	1,32	1,41	0,73
SK08-M	Suche	B12	**2,13**[b]	**1,75**	2,09	1,79	2,23	2,03	1,94	1,56
		B16	**0,43**[b]	**0,03**	0,45	0,006	0,59	0,27	0,31	−0,25
		B18	**1,29**[b]	**0,98**	1,35	0,92	1,45	1,14	1,25	0,71
SK09-M	Benutzerfreundlichkeit	B04	**1,57**[b]	**1,17**	1,60	1,14	1,71	1,43	1,49	0,85
		B06	**1,97**[b]	**1,67**	2,09	1,55	2,12	1,82	2,06	1,28
		B12	**3,07**[b]	**2,57**	3,09	2,55	3,23	2,91	2,94	2,19
		B16	**1,73**[b]	**1,37**	1,81	1,28	1,89	1,56	1,74	1,00
		B17	**2,35**[b]	**1,92**	2,38	1,89	2,44	2,26	2,32	1,52
		B17	2,37	2,01	2,40	1,98	**2,37**[b]	**2,37**	**2,43**	**1,59**
		B18	**2,04**[b]	**1,67**	2,12	1,58	2,18	1,90	2,07	1,26
SK11-M[d]	Eigenleistung	B06	0,93	0,82	**1,01**[a]	**0,74**	1,07	0,79	0,95	0,69
		B10	1,43	1,26	**1,44**[a]	**1,25**	1,52	1,33	1,35	1,17
		B17	0,49	0,36	**0,56**[a]	**0,29**	0,59	0,39	0,52	0,19
		B18	1,04	1,05	**1,18**[a]	**0,91**	1,23	0,85	1,12	0,98
SK12-F	Suchergebnis	B11	−0,89	−1,20	**−0,84**[a]	**−1,25**	−0,62	−1,16	−1,06	−1,34
		B16	**−3,62**[b]	**−4,14**	−3,56	−4,20	−3,41	−3,84	−3,71	−4,57
SK12-M	Suchergebnis	B12	**2,15**[b]	**1,76**	2,16	1,75	2,32	1,98	2,01	1,52
SK13-F	Benutzerfreundlichkeit	B01	0,23	−0,10	**0,22**[a]	**−0,09**	0,16	0,30	0,29	−0,48
		B17	−0,33	−0,41	−0,07	−0,66	**−0,34**[b]	**−0,31**	**0,20**	**−1,02**

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Entspricht auch der Skala SK18-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.63 (Fortsetzung)

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
SK14-F	Suche	B04	**−2,91**[b]	**−3,57**	−2,95	−3,53	−2,70	−3,12	−3,19	−3,94
		B05	**−1,64**[b]	**−2,09**	−1,71	−2,02	−1,63	−1,64	−1,78	−2,40
		B05	−1,98	−2,26	**−1,97**[a]	**−2,27**	−1,82	−2,14	−2,11	−2,40
		B06	**−2,43**[b]	**−2,96**	−2,39	−3,00	−2,24	−2,62	−2,54	−3,37
		B11	−1,58	−2,30	**−1,71**[a]	**−2,16**	−1,65	−1,51	−1,78	−2,82
		B12	**−1,09**[b]	**−1,88**	−1,20	−1,76	−0,98	−1,19	−1,42	−2,33
		B16	**−3,09**[b]	**−3,68**	−3,11	−3,66	−2,86	−3,31	−3,36	−4,00
		B18	**−2,13**[b]	**−2,67**	−2,11	−2,69	−1,94	−2,31	−2,29	−3,06
		M45	**0,11**[b]	**−0,35**	0,15	−0,40	0,30	−0,08	0,008	−0,71
		S05	**0,42**[b]	**−0,15**	0,50	−0,23	0,65	0,19	0,35	−0,65
SK14-M	Suche	B04	**1,17**[b]	**0,74**	1,18	0,73	1,34	1,01	1,03	0,46
		B05	**1,81**[b]	**1,45**	1,75	1,50	1,83	1,78	1,67	1,22
		B06	**1,26**[b]	**0,84**	1,27	0,83	1,39	1,14	1,15	0,53
		B10	**1,92**[b]	**1,48**	1,95	1,45	2,18	1,67	1,71	1,24
		B12	**2,34**[b]	**1,76**	2,26	1,84	2,47	2,21	2,05	1,48
		B16	**0,89**[b]	**0,49**	0,89	0,48	1,05	0,72	0,73	0,25
		B17	**1,73**[b]	**1,26**	1,71	1,29	1,83	1,64	1,59	0,93
		B17	1,34	1,01	**1,29**[a]	**1,05**	1,36	1,31	1,22	0,79
		B18	**1,53**[b]	**1,15**	1,55	1,12	1,67	1,38	1,43	0,87
SK15-F	Eigenleistung	B03	0,32	0,11	**0,72**[b]	**−0,29**	0,82	−0,18	0,63	−0,41
		B04	−3,14	−3,25	**−2,98**[b]	**−3,41**	−2,90	−3,37	−3,06	−3,45
		B05	−3,86	−4,00	**−3,65**[b]	**−4,22**	−3,63	−4,10	−3,67	−4,33
		M10	−0,64	−0,56	**−0,35**[b]	**−0,85**	−0,40	−0,87	−0,29	−0,83
		M16	−0,58	−0,52	**−0,30**[b]	**−0,80**	−0,31	−0,84	−0,28	−0,76
		M37	−0,50	−0,50	**−0,23**[b]	**−0,78**	−0,29	−0,71	−0,16	−0,85
		M45	−0,38	−0,51	**−0,20**[b]	**−0,69**	−0,16	−0,61	−0,25	−0,77
		S01	−0,007	−0,03	**0,27**[b]	**−0,31**	0,27	−0,28	0,28	−0,33
		S02	−0,02	−0,01	**0,29**[b]	**−0,32**	0,28	−0,32	0,30	−0,33
		S03	0,09	0,09	**0,34**[b]	**−0,16**	0,32	−0,13	0,37	−0,18
		S04	0,70	0,69	**1,00**[b]	**0,39**	0,95	0,44	1,04	0,33
		S05	0,49	0,58	**0,81**[b]	**0,26**	0,74	0,24	0,89	0,27
		S06	0,77	0,68	**1,01**[b]	**0,44**	0,96	0,58	1,06	0,29
SK16-F	Aufgabe	B01	−0,48	−0,73	**−0,44**[a]	**−0,77**	−0,18	−0,79	−0,70	−0,75
		B03	−0,25	−0,97	**−0,34**[a]	**−0,88**	0,11	−0,62	−0,79	−1,15
		B05	−2,37	−2,80	**−2,38**[a]	**−2,79**	−2,26	−2,48	−2,50	−3,10
		B06	−2,81	−3,02	**−2,75**[a]	**−3,07**	−2,56	−3,06	−2,95	−3,09
		B11	−1,75	−2,37	**−1,99**[a]	**−2,14**	−1,50	−2,00	−2,47	−2,28
		B12	**−1,52**[b]	**−2,15**	**−1,69**[a]	**−1,97**	−1,30	−1,74	−2,09	−2,21
		B18	−2,40	−2,58	**−2,29**[a]	**−2,69**	−2,08	−2,72	−2,50	−2,66
SK16-M	Aufgabe	B12	**2,82**[b]	**2,47**	2,78	2,51	3,02	2,62	2,55	2,39
		B16	**1,30**[b]	**1,04**	1,38	0,97	1,48	1,12	1,28	0,81

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Entspricht auch der Skala SK18-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Tab. E.63 (Fortsetzung)

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
SK17-F	Suche	B11	−1,90	−2,27	**−1,80**[a]	**−2,37**	−1,64	−2,15	−1,95	−2,59
		B16	**−3,62**[b]	**−4,15**	−3,57	−4,20	−3,40	−3,84	−3,74	−4,56
SK17-M	Suche	B06	**0,95**[b]	**0,64**	0,96	0,63	1,05	0,86	0,88	0,40
		B18	**1,26**[b]	**0,95**	1,31	0,89	1,40	1,12	1,23	0,67
SK18-F	Benutzerfreundlichkeit	B03	0,37	−0,55	**0,21**[a]	**−0,39**	0,38	0,37	0,04	−1,15
		B04	**−1,84**[b]	**−2,42**	−1,84	−2,42	−1,63	−2,05	−2,05	−2,79
		B05	**−0,36**[b]	**−1,13**	−0,53	−0,96	−0,33	−0,40	−0,73	−1,53
		B05	−0,89	−1,42	**−1,02**[a]	**−1,29**	−0,79	−0,99	−1,24	−1,59
		B06	**−1,92**[b]	**−2,53**	−1,95	−2,50	−1,75	−2,09	−2,15	−2,91
		B10	**−0,67**[b]	**−1,13**	−0,60	−1,20	−0,24	−1,10	−0,96	−1,30
		B11	−0,30	−1,23	**−0,66**[a]	**−0,88**	−0,14	−0,46	−1,18	−1,29
		B12	**−0,36**[b]	**−1,21**	−0,50	−1,07	−0,18	−0,53	−0,82	−1,61
		B16	**−2,04**[b]	**−2,61**	−2,02	−2,63	−1,81	−2,26	−2,22	−3,01
		B17	−1,31	−1,75	**−1,40**[a]	**−1,66**	−1,37	−1,25	−1,44	−2,06
		B18	**−1,29**[b]	**−1,90**	−1,28	−1,90	−1,07	−1,50	−1,49	−2,31
SK19-F	Eigenleistung	B05	−2,86	−2,99	**−2,72**[a]	**−3,13**	−2,64	−3,08	−2,79	−3,18
		B06	−2,76	−2,85	**−2,67**[a]	**−2,93**	−2,59	−2,93	−2,76	−2,94
		B10	−1,75	−1,99	**−1,75**[a]	**−1,99**	−1,62	−1,88	−1,87	−2,11
		B16	−2,48	−2,57	**−2,37**[a]	**−2,68**	−2,24	−2,73	−2,51	−2,63
		B17	−3,00	−3,10	**−2,86**[a]	**−3,23**	−2,88	−3,12	−2,85	−3,34
		B18	−1,84	−1,93	**−1,69**[a]	**−2,08**	−1,61	−2,06	−1,77	−2,09
SK-A	Accuracy (EUCS)	B01	2,50	2,05	**2,45**[a]	**2,10**	2,81	2,18	2,10	2,01
		B11	**1,63**[b]	**1,12**	1,68	1,07	1,74	1,52	1,62	0,62
		B12	**2,25**[b]	**1,72**	2,26	1,71	2,40	2,09	2,11	1,33
		B16	**−0,007**[b]	**−0,37**	0,08	−0,46	0,24	−0,25	−0,08	−0,66
		B17	**0,67**[b]	**0,38**	0,85	0,20	0,86	0,49	0,84	−0,09
		B18	**0,75**[b]	**0,40**	0,91	0,24	1,06	0,43	0,75	0,05
SK-C	Content (EUCS)	B05	**1,44**[b]	**1,07**	1,52	0,99	1,65	1,24	1,40	0,75
		B11	2,62	2,30	**2,58**[a]	**2,34**	2,74	2,50	2,42	2,18
SK-E[e]	Ease of Use (EUCS)	B01	4,44	4,43	**4,58**[a]	**4,29**	4,54	4,34	4,62	4,24
		B17	3,63	3,56	3,79	3,40	**3,64**[b]	**3,62**	**3,95**	**3,18**
SK-E-09	EUCS-Skala-2009	B06	**1,13**[b]	**0,87**	1,23	0,77	1,28	0,98	1,18	0,57
		B12	**2,04**[b]	**1,57**	2,04	1,58	2,20	1,88	1,87	1,27
		B16	**0,67**[b]	**0,32**	0,72	0,27	0,82	0,51	0,61	0,02
SK-E-13	EUCS-Skala-2013	B01	3,16	2,78	**3,09**[a]	**2,86**	3,23	3,10	2,95	2,62
		B06	**1,32**[b]	**1,05**	1,39	0,99	1,44	1,20	1,33	0,78
		B12	**2,76**[b]	**2,33**	2,73	2,36	2,86	2,65	2,59	2,06
		B16	**0,96**[b]	**0,65**	1,05	0,56	1,13	0,79	0,96	0,34

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Entspricht auch der Skala SK18-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.63 (Fortsetzung)

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
SK-E-88	EUCS-Skala-1988	B04	**1,26**[b]	**0,99**	1,33	0,91	1,42	1,10	1,25	0,72
		B12	**2,47**[b]	**2,15**	2,51	2,11	2,65	2,29	2,37	1,92
		B16	**1,26**[b]	**1,02**	1,36	0,92	1,44	1,09	1,28	0,76
SK-G-13	Gesamtskala-2013	B05	**1,22**[b]	**0,81**	1,22	0,81	1,33	1,11	1,11	0,50
		B06	**1,22**[b]	**0,94**	1,25	0,91	1,33	1,11	1,17	0,72
		B11	1,39	1,11	**1,44**[a]	**1,06**	1,59	1,18	1,29	0,94
		B12	**2,17**[b]	**1,73**	2,10	1,80	2,28	2,06	1,92	1,53
		B16	**0,79**[b]	**0,47**	0,83	0,43	0,96	0,61	0,71	0,24
		B17	**1,42**[b]	**1,02**	1,44	1,00	1,52	1,32	1,37	0,68
SK-T	Timeliness (EUCS)	B03	3,53	3,22	**3,59**[a]	**3,16**	3,72	3,34	3,46	2,99
		B05	2,34	2,34	**2,49**[a]	**2,19**	2,50	2,17	2,47	2,21
		B11	3,83	3,79	**3,83**[a]	**3,79**	3,91	3,75	3,75	3,83
		B12	3,34	3,00	**3,24**[a]	**3,10**	3,42	3,26	3,06	2,94
		B17	2,35	2,27	**2,47**[a]	**2,15**	2,42	2,28	2,52	2,01
E01	Ich glaube, ich werde in zehn Minuten ... relevante Dokumente finden.	B18	0,29	−0,07	**0,59**[b]	**−0,37**	0,94	−0,35	0,25	−0,39
		S01	4,24	4,74	**4,93**[b]	**4,05**	4,86	3,62	5,00	4,48
		S05	5,66	5,83	**6,17**[b]	**5,33**	6,11	5,22	6,22	5,44
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	B06	**1,28**[b]	**0,89**	1,28	0,88	1,47	1,08	1,09	0,68
		B11	3,09	2,50	**2,83**[a]	**2,76**	3,19	2,99	2,47	2,53
		B12	**2,31**[b]	**1,79**	2,19	1,91	2,42	2,20	1,97	1,62
		B12	2,25	1,81	**2,12**[a]	**1,94**	2,36	2,14	1,89	1,73
		B16	**1,21**[b]	**0,86**	1,26	0,81	1,43	0,99	1,09	0,63
		B18	**1,36**[b]	**0,97**	1,39	0,95	1,57	1,15	1,20	0,74
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	B11	3,33	2,92	**3,33**[a]	**2,92**	3,33	3,33	3,33	2,50
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	B01	3,88	3,69	**3,91**[a]	**3,66**	4,12	3,64	3,69	3,69
		B03	3,67	3,00	**3,50**[a]	**3,17**	4,00	3,33	3,00	3,00
		B04	**1,25**[b]	**0,92**	1,25	0,91	1,35	1,14	1,16	0,68
		B05	**2,10**[b]	**1,66**	2,06	1,70	2,23	1,98	1,89	1,42
		B11	**2,57**[b]	**1,75**	2,32	1,99	2,59	2,54	2,04	1,45
		B11	1,87	1,35	**1,68**[a]	**1,54**	1,86	1,88	1,49	1,20
		B16	**0,92**[b]	**0,59**	0,92	0,59	1,04	0,79	0,81	0,38
		B17	1,39	1,10	**1,30**[a]	**1,19**	1,38	1,39	1,22	0,98

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Entspricht auch der Skala SK18-M.

[d] Entspricht auch den Skalen SK15-M und SK19-M.

[e] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.63 (Fortsetzung)

ID	Beschreibung	Kov.	S_G	S_S	E_H	E_N	$I_{G,H}$	$I_{G,N}$	$I_{S,H}$	$I_{S,N}$
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	B01	**3,33**[a]	**3,00**	3,50	2,83	3,67	3,00	3,33	2,67
		B01	3,50	2,67	**3,33**[a]	**2,83**	3,67	3,33	3,00	2,33
		B11	**1,21**[a]	**0,66**	1,19	0,68	1,27	1,14	1,10	0,22
		B11	1,99	1,42	**1,79**[a]	**1,62**	2,06	1,92	1,52	1,32
		M16	**2,76**[a]	**2,51**	2,91	2,36	2,96	2,57	2,87	2,15
E06-M	Erwartungsskala	B04	**1,15**[b]	**0,77**	1,19	0,73	1,35	0,95	1,04	0,50
		B06	**1,74**[b]	**1,34**	1,76	1,32	1,86	1,61	1,65	1,03
		B10	**2,12**[b]	**1,69**	2,18	1,64	2,38	1,86	1,97	1,41
		B11	3,40	3,11	**3,38**[a]	**3,14**	3,57	3,23	3,18	3,05
		B16	**1,01**[b]	**0,63**	1,07	0,57	1,22	0,80	0,93	0,34
		B18	**1,88**[b]	**1,48**	1,96	1,41	2,08	1,68	1,84	1,13
		M45	**3,08**[b]	**2,80**	3,19	2,69	3,28	2,87	3,10	2,51

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Entspricht auch der Skala SK18-M.
[d] Entspricht auch den Skalen SK15-M und SK19-M.
[e] Entspricht auch der Skala SK13-M.

E.6.3. Teststatistiken der Kovarianzanalyse

Die in diesem Abschnitt dargestellten Tabellen enthalten die Teststatistiken zu den im Rahmen des dritten Experiments durchgeführten Kovarianzanalysen. Um die entsprechenden Signifikanzniveaus zu den einzelnen Variablen-Kovariaten-Kombinationen leichter auffindbar zu machen, sind auch die Teststatistiken in drei Tabellen getrennt nach Benutzerleistung und Benutzerzufriedenheit sowie den beiden im Kontext der Zufriedenheitsanalyse betrachteten Kovariatengruppen aufgeführt.

Tab. E.64.: Teststatistik der Kovarianzanalyse zur Untersuchung des Einflusses demographischer und erfahrungsbezogener Störfaktoren auf die Benutzerleistung in Stichprobe SP_A. In allen Fällen handelt es sich um robuste Kovarianzanalysen.

				Kovariate		System		Erwartung		Interaktion	
ID	Beschreibung	Kov.	n	F	p	F	p	F	p	F	p
M02	Anz. aufg. Dok. (erste 10 Dok.)	K02	116	0,02	0,88	0,91	0,34	0,97	0,33	2,02	**0,16**[a]
		K03	116	0,42	0,52	0,73	0,39	$-2{\cdot}10^{-5}$	1,00	2,78	**0,10**[a]
		K04	116	0,01	0,92	1,65	0,20	0,02	0,89	2,57	**0,11**[a]
		K08	116	0,03	0,85	0,93	0,34	0,97	0,33	2,01	**0,16**[a]
		K09	116	0,008	0,93	2,28	0,13	0,02	0,88	3,56	**0,06**[a]
M11	Anz. rel. bew. Dok. (erste 10 Dok.)	K03	116	1,88	0,17	1,63	0,20	2,25	0,14	5,41	**0,02**[b,d]
		K09	116	1,05	0,31	1,75	0,19	2,28	0,13	5,03	**0,03**[b,d]
M17	Anz. richtig rel. bew. Dok. (erste 10 Dok.)	K02	116	0,46	0,50	3,65	0,06	1,72	0,19	8,51	**0,004**[b,d]
		K08	116	−0,03	1,00	5,90	**0,02**[b,d]	2,03	0,16	4,00	0,05
M21	Anz. aufg. eher rel. Dok.	K09	116	0,37	0,55	9,51	**0,003**[b,d]	0,008	0,93	0,21	0,65
		K10	116	6,04	**0,02**[d]	10,35	**0,002**[b,c]	0,07	0,79	0,02	0,88

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
Fortsetzung auf nächster Seite
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

Tab. E.64 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
M26	Anz. falsch eher irrel. bew. Dok.	K02	116	0,41	0,53	2,48	**0,12**[a]	$-2{\cdot}10^{-5}$	1,00	0,43	0,51
		K05	116	1,68	0,20	2,94	**0,09**[a]	0,49	0,49	4,51	0,04
		K08	116	1,56	0,21	2,44	**0,12**[a]	0,04	0,84	0,37	0,54
M29	Anz. falsch eher irrel. bew. rel. Dok.	K03	116	0,02	0,88	0,36	0,55	8,32	**0,005**[b,d]	0,36	0,55
		K04	116	0,03	0,87	0,36	0,55	8,32	**0,005**[b,d]	0,36	0,55
		K09	116	0,03	0,87	0,36	0,55	8,32	**0,005**[b,d]	0,36	0,55
		K10	116	1,05	0,31	0,69	0,41	6,32	**0,01**[b,d]	0,12	0,73
B06	Durchschn. Bew. rel. Dok. (letzte Suche)	K10	96	0,62	0,43	2,65	**0,11**[a]	0,55	0,46	1,62	0,21
Z02	Durchschn. Betrachtungsz. falsch bew. Dok.	K05	80	0,45	0,51	1,96	**0,17**[a]	0,23	0,63	0,64	0,43
		K08	80	1,80	0,18	1,97	**0,16**[a]	0,71	0,40	1,04	0,31
		K10	80	0,83	0,37	1,65	**0,20**[a]	0,20	0,66	0,94	0,33
Z05-log	Durchschn. Betrachtungsz. irrel. bew. Dok.	K02	88	1,97	0,16	2,82	**0,10**[a]	0,12	0,74	0,98	0,33
		K10	88	1,23	0,27	3,01	**0,09**[a]	0,38	0,54	1,53	0,22
V14	$\frac{\text{Anz. richtig irrel. bew. Dok.}}{\text{Anz. aufg. Dok.}}$	K04	116	1,15	0,29	25,65	$2{\cdot}10^{-6}$	2,44	**0,12**[a]	0,05	0,82
V29	$\frac{\text{Anz. richtig rel. bew. Dok.}}{\text{Anz. aufg. rel. Dok.}}$	K02	108	0,77	0,38	4,93	**0,03**[b,d]	0,002	0,97	0,01	0,91
		K03	108	$-3{\cdot}10^{-4}$	1,00	5,06	**0,03**[b,d]	0,02	0,89	0,10	0,76
		K04	108	1,14	0,29	5,57	**0,02**[b,d]	0,01	0,91	0,04	0,85
		K05	108	5,17	**0,03**[d]	7,06	**0,009**[b,d]	0,26	0,61	0,18	0,68
		K08	108	2,07	0,15	5,16	**0,03**[b,d]	$3{\cdot}10^{-4}$	0,99	0,11	0,74
		K09	108	1,69	0,20	6,04	**0,02**[b,d]	0,01	0,91	0,05	0,83
		K10	108	6,38	**0,01**[c]	4,93	**0,03**[b,d]	0,31	0,58	0,18	0,67
V34	$\frac{\text{Anz. aufg. eher rel. Dok.}}{\text{Anz. eher rel. Dok. im Korpus}}$	K05	116	0,61	0,43	6,47	**0,01**[b,d]	0,02	0,88	1,33	0,25
		K09	116	0,04	0,85	6,84	**0,01**[b,d]	0,26	0,61	0,61	0,44
		K10	116	4,09	**0,05**[d]	8,32	**0,005**[b,c]	0,07	0,79	0,21	0,65
V58	$\frac{\text{Anz. richtig eher rel. bew. Dok.}}{\text{Anz. eher rel. bew. Dok.}}$	K02	68	0,08	0,78	0,84	0,36	2,13	**0,15**[a]	0,87	0,35

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

Tab. E.65.: Teststatistik der Kovarianzanalyse zur Untersuchung des Einflusses demographischer und erfahrungsbezogener Störfaktoren auf die Benutzerzufriedenheit in Stichprobe SP_A.

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
F11	Liefert die Suchmaschine aktuelle Information?	K03	116	0,93	0,34	0,14	0,71	6,57	**0,01**[b,b]	0,14	0,71

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

Fortsetzung auf nächster Seite

Tab. E.65 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	K02	116	0,43	0,51	2,43	**0,12**[a]	10,01	0,002	0,33	0,57
SK03-M	Benutzerfreundlichkeit	K05	116	2,16	0,14	2,80	0,10	16,17	$1{\cdot}10^{-4}$	6,44	**0,01**[b,b]
SK14-M	Suche	K02	116	$3{\cdot}10^{-4}$	0,99	6,17	**0,01**[b,b]	14,33	$2{\cdot}10^{-4}$	0,93	0,34
		K05	116	$3{\cdot}10^{-4}$	0,99	7,83	**0,006**[b,b]	20,79	$1{\cdot}10^{-5}$	2,69	0,10
SK15-F	Eigenleistung	K02	116	0,003	0,96	0,03	0,87	8,08	**0,005**[b,b]	0,11	0,74
		K03	116	1,23	0,27	0,01	0,91	8,40	**0,005**[b,b]	0,31	0,58
		K05	116	0,14	0,71	0,05	0,83	7,61	**0,007**[b,b]	0,17	0,68
		K09	116	0,33	0,57	0,04	0,84	7,78	**0,006**[b,b]	0,09	0,76
		K10	116	0,009	0,93	0,03	0,86	7,95	**0,006**[b,b]	0,13	0,72
E01	Ich glaube, ich werde in zehn Minuten ... relevante Dokumente finden.	K04	116	$1{\cdot}10^{-5}$	1,00	2,91	0,09	9,91	**0,002**[b,b]	0,83	0,36
		K10	116	4,18	**0,04**[b]	1,79	0,18	10,37	**0,002**[b,b]	0,37	0,55

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

Tab. E.66.: Teststatistik der Kovarianzanalyse zur Untersuchung des Einflusses leistungsbezogener Störfaktoren auf die Benutzerzufriedenheit in Stichprobe SP_A.

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
F01	Liefert die Suchmaschine genau die Information, die Sie benötigen?	B11	24	0,85	0,37	0,61	0,44	1,53	**0,23**[a]	1,48	0,24
		B17	60	29,53	**$1{\cdot}10^{-6}$**[c]	8,67	**0,005**[b,d]	18,13	$8{\cdot}10^{-5}$ [c]	4,06	0,05
F02	Befriedigen die gelieferten Inhalte Ihr Informationsbedürfnis?	B01	40	0,65	0,42	2,00	0,17	0,83	**0,37**[a]	0,28	0,60
		B04	116	71,98	**$1{\cdot}10^{-13}$**[c]	8,19	**0,005**[b,c]	10,05	0,002	$4{\cdot}10^{-4}$	0,98
		B05	72	30,65	**$6{\cdot}10^{-7}$**[c]	11,23	**0,001**[b,c]	10,39	0,002	0,86	0,36
		B16	116	74,03	**$6{\cdot}10^{-14}$**[c]	7,00	**0,009**[b,c]	9,16	0,003 [c]	0,06	0,80
		B17	60	20,88	**$3{\cdot}10^{-5}$**[c]	5,66	**0,02**[b,d]	8,82	0,004 [c]	0,62	0,43
F03	Liefert die Suchmaschine Dokumente, die genau das beinhalten, was Sie benötigen?	B04	116	56,81	**$1{\cdot}10^{-11}$**[c]	9,79	**0,002**[b,c]	11,14	0,001	3,08	0,08
		B10	76	28,12	**$1{\cdot}10^{-6}$**[c]	1,97	0,16	2,16	**0,15**[a]	0,02	0,89
		B11	24	11,37	**0,003**[c]	$4{\cdot}10^{-7}$	1,00	1,66	**0,21**[a]	0,83	0,37
		B16	116	46,42	**$5{\cdot}10^{-10}$**[c]	7,83	**0,006**[b,d]	12,01	$8{\cdot}10^{-4}$ [c]	1,82	0,18
F04	Liefert die Suchmaschine genügend Information?	B10	76	4,44	**0,04**[c]	0,41	0,52	3,56	**0,06**[a]	0,02	0,88
		B11	24	10,54	**0,004**[d]	3,33	0,08	0,28	**0,60**[a]	0,09	0,77

Fortsetzung auf nächster Seite

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.
[e] Entspricht auch der Skala SK18-M.
[f] Entspricht auch den Skalen SK15-M und SK19-M.
[g] Entspricht auch der Skala SK13-M.

Tab. E.66 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
F05	Ist die Suchmaschine präzise?	B04	116	39,96	$\mathbf{6{\cdot}10^{-9}}$[c]	11,86	$\mathbf{8{\cdot}10^{-4}}$[b,c]	29,57	$3{\cdot}10^{-7}$	0,65	0,42
		B06	96	28,91	$\mathbf{6{\cdot}10^{-7}}$[c]	11,33	**0,001**[b,d]	17,08	$8{\cdot}10^{-5}$	1,29	0,26
		B12	56	7,57	**0,008**[c]	9,76	**0,003**[b,d]	7,61	0,008	3,33	0,07
		B16	116	52,74	$\mathbf{6{\cdot}10^{-11}}$[c]	14,70	$\mathbf{2{\cdot}10^{-4}}$[b,c]	26,51	$1{\cdot}10^{-6}$	0,06	0,80
F06	Sind Sie mit der Genauigkeit der Suchmaschine zufrieden?	B12	56	11,37	**0,001**[c]	8,47	**0,005**[b,d]	8,48	0,005	3,44	0,07
		B16	116	87,95	$\mathbf{9{\cdot}10^{-16}}$[c]	13,23	$\mathbf{4{\cdot}10^{-4}}$[b,c]	36,59	$2{\cdot}10^{-8}$ [c]	2,48	0,12
F07	Finden Sie die Präsentation der Ergebnisse hilfreich?	B05	72	8,47	**0,005**[c]	0,94	0,34	2,09	**0,15**[a]	1,70	0,20
		B10	76	5,12	**0,03**[c]	0,47	0,49	3,19	**0,08**[a]	1,35	0,25
		B11	24	2,08	0,17	2,35	0,14	0,45	**0,51**[a]	1,22	0,28
F08	Ist die Suchmaschine benutzerfreundlich?	S06	116	$2{\cdot}10^{-4}$	0,99	3,46	0,07	37,14	$2{\cdot}10^{-8}$ [c]	7,37	**0,008**[b,d]
F10	Bekommen Sie die Information, die Sie benötigen rechtzeitig?	B10	76	0,94	0,34	0,01	0,91	1,64	**0,20**[a]	0,22	0,64
		B11	24	0,00	1,00	$1{\cdot}10^{-14}$	1,00	0,00	**1,00**[a]	$7{\cdot}10^{-15}$	1,00
		B12	56	2,16	0,15	1,31	0,26	0,93	**0,34**[a]	0,03	0,85
F12	Ist die Suchmaschine erfolgreich?	M10	116	16,56	$\mathbf{9{\cdot}10^{-5}}$[c]	0,23	0,63	24,77	$2{\cdot}10^{-6}$	7,77	**0,006**[b,c]
		M37	116	13,80	$\mathbf{3{\cdot}10^{-4}}$[c]	0,62	0,43	30,08	$3{\cdot}10^{-7}$ [c]	7,43	**0,007**[b,d]
F13	Sind Sie mit der Suchmaschine zufrieden?	B06	96	39,00	$\mathbf{1{\cdot}10^{-8}}$[c]	6,12	**0,02**[b,d]	12,29	$7{\cdot}10^{-4}$	1,66	0,20
		B11	24	0,06	0,81	0,18	0,68	1,50	**0,24**[a]	0,07	0,80
		B16	116	76,82	$\mathbf{2{\cdot}10^{-14}}$[c]	6,99	**0,009**[b,c]	28,18	$6{\cdot}10^{-7}$ [c]	2,41	0,12
F14	Es war einfach, die Aufgabe zu bearbeiten.	B06	96	13,85	$\mathbf{3{\cdot}10^{-4}}$[c]	2,52	0,12	1,28	**0,26**[a]	0,83	0,36
		B11	24	10,95	**0,004**[c]	0,48	0,50	0,02	**0,89**[a]	3,68	0,07
		B12	56	10,63	**0,002**[c]	4,31	0,04	1,60	**0,21**[a]	2,20	0,14
		B17	60	14,12	$\mathbf{4{\cdot}10^{-4}}$[c]	0,99	0,32	1,42	**0,24**[a]	0,13	0,72
		B18	84	15,53	$\mathbf{2{\cdot}10^{-4}}$[c]	3,78	0,06	2,81	**0,10**[a]	1,01	0,32
F16	Es war einfach, mit der Suchmaschine relevante Dokumente zu finden.	B05	72	11,49	**0,001**[c]	6,40	**0,01**[b,d]	10,17	0,002	3,02	0,09
		B11	24	0,44	0,51	0,02	0,88	0,02	**0,90**[a]	0,59	0,45
F17	Es war nachvollziehbar, warum welche Dokumente gefunden wurden.	B05	72	5,67	**0,02**[d]	3,47	0,07	0,32	**0,57**[a]	1,44	0,23
		B06	96	17,05	$\mathbf{8{\cdot}10^{-5}}$[c]	6,75	**0,01**[b,d]	4,04	0,05	0,29	0,59
		B10	76	5,22	**0,03**[d]	3,55	0,06	1,38	**0,24**[a]	0,12	0,73
		B12	56	2,65	0,11	11,32	**0,001**[b,c]	5,64	0,02	1,27	0,27
		B12	56	2,27	0,14	6,21	0,02	2,98	**0,09**[a]	0,11	0,75
		B16	116	27,38	$\mathbf{8{\cdot}10^{-7}}$[c]	8,27	**0,005**[b,c]	10,50	0,002	0,86	0,36
		B17	60	4,79	**0,03**[c]	2,57	0,11	2,08	**0,15**[a]	1,27	0,27
		B18	84	13,65	$\mathbf{4{\cdot}10^{-4}}$[c]	3,64	0,06	3,87	**0,05**[a]	0,60	0,44

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.

[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

[e] Entspricht auch der Skala SK18-M.

[f] Entspricht auch den Skalen SK15-M und SK19-M.

[g] Entspricht auch der Skala SK13-M.

Tab. E.66 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
F18	Die meisten Dokumente waren in Bezug auf meine Suchanfragen relevant.	B04	116	39,25	**$7{\cdot}10^{-9}$**[c]	14,44	**$2{\cdot}10^{-4}$**[b,c]	8,88	0,004	0,98	0,32
		B05	72	28,80	**$1{\cdot}10^{-6}$**[c]	11,67	**0,001**[b,c]	6,05	0,02	0,02	0,90
		B06	96	26,99	**$1{\cdot}10^{-6}$**[c]	6,09	**0,02**[b,d]	5,41	0,02	0,18	0,67
		B12	56	17,01	**$1{\cdot}10^{-4}$**[c]	7,96	**0,007**[b,c]	3,87	0,05	0,62	0,44
		B16	116	60,31	**$4{\cdot}10^{-12}$**[c]	16,23	**$1{\cdot}10^{-4}$**[b,c]	8,94	0,003	0,45	0,50
		B18	84	26,14	**$2{\cdot}10^{-6}$**[c]	7,10	**0,009**[b,d]	4,59	0,04	1,34	0,25
		B18	84	25,76	**$3{\cdot}10^{-6}$**[c]	3,34	0,07	3,04	**0,09**[a]	2,57	0,11
		S06	116	7,96	**0,006**[c]	9,35	**0,003**[b,d]	14,06	$3{\cdot}10^{-4}$	0,005	0,94
F19	Die Suchmaschine hat die Suchergebnisse gut gefiltert.	B05	72	13,00	**$6{\cdot}10^{-4}$**[c]	2,58	**0,11**[a]	5,48	0,02	0,09	0,76
		B10	76	18,85	**$5{\cdot}10^{-5}$**[c]	2,94	**0,09**[a]	8,84	0,004	0,02	0,90
		B12	56	4,89	**0,03**[c]	2,24	**0,14**[a]	4,43	0,04	0,85	0,36
		B12	56	4,38	**0,04**[c]	4,13	0,05	3,27	**0,08**[a]	2,17	0,15
		B17	60	9,29	**0,004**[c]	3,96	0,05	3,03	**0,09**[a]	0,94	0,34
		S03	116	1,52	0,22	1,90	**0,17**[a]	12,16	$7{\cdot}10^{-4}$	0,83	0,36
		S04	116	0,98	0,33	2,38	**0,13**[a]	13,03	$5{\cdot}10^{-4}$	0,21	0,65
		S06	116	3,33	0,07	3,30	**0,07**[a]	15,34	$2{\cdot}10^{-4}$	1,60	0,21
F20	Ich bin davon überzeugt, dass ich alle für das Thema relevanten Dokumente gefunden habe.	B05	72	17,15	**$1{\cdot}10^{-4}$**[c]	0,30	0,59	2,82	**0,10**[a]	0,66	0,42
		B10	76	14,81	**$3{\cdot}10^{-4}$**[c]	0,12	0,73	0,82	**0,37**[a]	1,41	0,24
		B17	60	14,91	**$3{\cdot}10^{-4}$**[c]	0,14	0,71	0,88	**0,35**[a]	0,17	0,68
		B18	84	7,88	**0,006**[c]	$6{\cdot}10^{-5}$	0,99	1,46	**0,23**[a]	0,25	0,62
F22	Ich bin mit den Suchergebnissen zufrieden.	B04	116	59,54	**$6{\cdot}10^{-12}$**[c]	6,95	**0,01**[b,c]	13,39	$4{\cdot}10^{-4}$ [c]	0,23	0,63
		B05	72	20,32	**$3{\cdot}10^{-5}$**[c]	9,33	**0,003**[b,d]	8,01	0,006 [c]	1,50	0,22
		B11	24	4,18	0,06	7,75	**0,01**[b,d]	3,25	0,09	2,12	0,16
		B11	24	1,64	0,22	0,52	0,48	1,81	**0,19**[a]	0,82	0,38
		B12	56	8,16	**0,006**[c]	8,37	**0,006**[b,c]	5,77	0,02	0,39	0,53
		B16	116	53,69	**$4{\cdot}10^{-11}$**[c]	9,67	**0,002**[b,d]	9,94	0,002	0,95	0,33
F23	Ich bin mit meiner Suchleistung zufrieden.	B05	72	14,48	**$3{\cdot}10^{-4}$**[c]	0,43	0,52	2,68	**0,11**[a]	0,26	0,61
		B06	96	15,73	**$1{\cdot}10^{-4}$**[c]	0,07	0,79	1,46	**0,23**[a]	0,05	0,82
		B10	76	10,47	**0,002**[c]	0,66	0,42	1,02	**0,31**[a]	0,10	0,75
		B11	24	0,14	0,71	0,44	0,51	2,61	**0,12**[a]	0,21	0,65
		B16	116	23,59	**$4{\cdot}10^{-6}$**[c]	1,07	0,30	2,91	**0,09**[a]	0,007	0,94
		B17	60	19,62	**$5{\cdot}10^{-5}$**[c]	0,03	0,86	2,50	**0,12**[a]	1,47	0,23
		B18	84	14,33	**$3{\cdot}10^{-4}$**[c]	0,02	0,90	3,52	**0,06**[a]	0,15	0,70
		M10	116	15,41	**$2{\cdot}10^{-4}$**[c]	0,30	0,59	2,61	**0,11**[a]	0,006	0,94
		M45	116	10,25	**0,002**[c]	0,58	0,45	2,28	**0,13**[a]	0,17	0,68

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.

[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

[e] Entspricht auch der Skala SK18-M.

[f] Entspricht auch den Skalen SK15-M und SK19-M.

[g] Entspricht auch der Skala SK13-M.

Tab. E.66 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
F24	Es hat Spaß gemacht, mit der Suchmaschine zu arbeiten.	B05	72	7,87	**0,007**[d]	16,19	**1·10^{-4}**[b,d]	12,32	8·10^{-4}	0,70	0,41
		B05	72	1,24	0,27	4,79	0,03	2,75	**0,10**[a]	0,16	0,69
		B11	24	11,90	**0,003**[d]	3,17	0,09	1,70	**0,21**[a]	0,05	0,82
		B12	56	1,28	0,26	14,71	**3·10^{-4}**[b,c]	10,22	0,002	0,76	0,39
		B16	116	18,89	**3·10^{-5}**[c]	6,68	**0,01**[b,c]	5,97	0,02 [c]	0,20	0,65
		B17	60	7,18	**0,01**[d]	9,96	**0,003**[b,d]	8,51	0,005	3,85	0,05
		B18	84	10,58	**0,002**[c]	7,51	**0,008**[b,d]	7,66	0,007	0,12	0,73
F25	Insgesamt bin ich mit der Suchmaschine sehr zufrieden.	B01	40	2,56	0,12	2,65	0,11	1,53	**0,22**[a]	1,38	0,25
		B11	24	4,17	0,06	2,07	0,17	1,65	**0,21**[a]	1,63	0,22
		B16	116	57,38	**1·10^{-11}**[c]	7,72	**0,006**[b,c]	18,69	3·10^{-5}	2,64	0,11
		B17	60	15,44	**2·10^{-4}**[c]	4,37	0,04	12,43	9·10^{-4}	6,44	**0,01**[b,d]
F26	Ich könnte mir vorstellen, diese Suchmaschine als Standardsuchmaschine in meinem Browser einzustellen.	B03	28	1,61	0,22	6,47	0,02	2,17	**0,15**[a]	0,13	0,72
		B16	116	29,27	**4·10^{-7}**[c]	6,05	**0,02**[b,c]	13,69	3·10^{-4}	2,36	0,13
		B17	60	12,55	**8·10^{-4}**[c]	3,00	0,09	2,73	**0,10**[a]	2,90	0,09
SK-A	Accuracy (EUCS)	B01	40	5,72	**0,02**[d]	3,83	0,06	3,27	**0,08**[a]	1,59	0,22
		B11	24	17,55	**5·10^{-4}**[d]	6,91	**0,02**[b,d]	11,70	0,003 [d]	4,22	0,05
		B12	56	9,77	**0,003**[c]	8,35	**0,006**[b,d]	10,87	0,002	2,10	0,15
		B16	116	85,99	**2·10^{-15}**[c]	13,54	**4·10^{-4}**[b,c]	28,06	6·10^{-7}	0,19	0,66
		B17	60	34,40	**3·10^{-7}**[c]	6,48	**0,01**[b,d]	23,95	9·10^{-6} [c]	5,50	0,02
		B18	84	30,52	**4·10^{-7}**[c]	6,04	**0,02**[b,c]	20,43	2·10^{-5}	0,10	0,76
SK-C	Content (EUCS)	B05	72	26,80	**2·10^{-6}**[c]	9,47	**0,003**[b,c]	19,19	4·10^{-5} [c]	1,01	0,32
		B11	24	4,64	**0,04**[d]	2,70	0,12	1,62	**0,22**[a]	2·10^{-4}	0,99
SK-E-09	EUCS-Skala-2009	B06	96	29,36	**5·10^{-7}**[c]	5,04	**0,03**[b,d]	16,52	1·10^{-4} [c]	1,97	0,16
		B12	56	15,09	**3·10^{-4}**[c]	9,80	**0,003**[b,c]	11,78	0,001 [c]	0,95	0,33
		B16	116	73,36	**7·10^{-14}**[c]	13,13	**4·10^{-4}**[b,c]	22,51	6·10^{-6}	2,01	0,16
SK-E-13	EUCS-Skala-2013	B01	40	8,83	**0,005**[d]	3,54	0,07	2,41	**0,13**[a]	0,51	0,48
		B06	96	36,87	**3·10^{-8}**[c]	6,57	**0,01**[b,c]	13,63	4·10^{-4}	2,32	0,13
		B12	56	10,82	**0,002**[c]	12,07	**0,001**[b,d]	11,36	0,001	2,15	0,15
		B16	116	67,59	**4·10^{-13}**[c]	10,68	**0,001**[b,c]	25,92	1·10^{-6} [c]	2,65	0,11
SK-E-88	EUCS-Skala-1988	B04	116	55,72	**2·10^{-11}**[c]	10,77	**0,001**[b,c]	27,68	7·10^{-7} [c]	1,88	0,17
		B12	56	11,22	**0,002**[c]	6,03	**0,02**[b,d]	9,75	0,003	0,07	0,79
		B16	116	60,82	**4·10^{-12}**[c]	7,70	**0,006**[b,c]	24,27	3·10^{-6}	1,17	0,28

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.

[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

[e] Entspricht auch der Skala SK18-M.

[f] Entspricht auch den Skalen SK15-M und SK19-M.

[g] Entspricht auch der Skala SK13-M.

Tab. E.66 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
SK-G-13	Gesamtskala-2013	B05	72	39,61	**$3{\cdot}10^{-8}$**[c]	14,23	**$3{\cdot}10^{-4}$**[b,c]	15,34	$2{\cdot}10^{-4}$ [c]	3,58	0,06
		B06	96	33,74	**$9{\cdot}10^{-8}$**[c]	6,46	**0,01**[b,d]	9,12	0,003	1,09	0,30
		B11	24	15,21	**0,001**[d]	1,54	0,23	2,46	**0,13**[a]	0,002	0,96
		B12	56	16,83	**$1{\cdot}10^{-4}$**[c]	12,09	**0,001**[b,c]	5,71	0,02	0,48	0,49
		B16	116	68,04	**$4{\cdot}10^{-13}$**[c]	11,34	**0,001**[b,c]	18,62	$3{\cdot}10^{-5}$	0,48	0,49
		B17	60	33,53	**$3{\cdot}10^{-7}$**[c]	10,78	**0,002**[b,d]	13,70	$5{\cdot}10^{-4}$ [c]	3,78	0,06
SK-T	Timeliness (EUCS)	B03	28	0,16	0,69	0,78	0,39	1,81	**0,19**[a]	0,02	0,90
		B05	72	6,69	**0,01**[d]	0,001	0,97	3,13	**0,08**[a]	0,06	0,81
		B11	24	0,04	0,84	0,04	0,85	0,03	**0,87**[a]	0,25	0,62
		B12	56	1,16	0,29	3,47	0,07	0,62	**0,43**[a]	0,02	0,89
		B17	60	6,68	**0,01**[d]	0,16	0,69	3,63	**0,06**[a]	1,32	0,26
SK01-M	Genauigkeit	B11	24	4,98	**0,04**[d]	0,37	0,55	2,11	**0,16**[a]	0,14	0,71
		B16	116	78,12	**$2{\cdot}10^{-14}$**[c]	14,97	**$2{\cdot}10^{-4}$**[b,c]	29,68	$3{\cdot}10^{-7}$	0,67	0,42
SK02-M	Inhalt	B01	40	0,85	0,36	$3{\cdot}10^{-4}$	0,99	1,77	**0,19**[a]	2,95	0,09
		B05	72	18,97	**$5{\cdot}10^{-5}$**[c]	6,53	**0,01**[b,d]	14,85	$3{\cdot}10^{-4}$ [c]	0,81	0,37
		B11	24	2,38	0,14	0,04	0,84	0,03	**0,87**[a]	0,01	0,92
SK03-M	Benutzerfreundlichkeit	B04	116	19,69	**$2{\cdot}10^{-5}$**[c]	9,59	**0,002**[b,c]	18,43	$4{\cdot}10^{-5}$	4,95	0,03
		B16	116	24,88	**$2{\cdot}10^{-6}$**[c]	6,80	**0,01**[b,c]	20,72	$1{\cdot}10^{-5}$ [c]	4,03	0,05
		B17	60	6,53	**0,01**[d]	4,74	0,03	10,92	0,002 [d]	12,10	**0,001**[b,c]
		B18	84	17,13	**$9{\cdot}10^{-5}$**[c]	5,78	**0,02**[b,d]	16,64	$1{\cdot}10^{-4}$	6,69	**0,01**[b,d]
		M45	116	5,01	**0,03**[c]	5,23	**0,02**[b,d]	12,71	$5{\cdot}10^{-4}$	6,98	**0,009**[b,c]
		S03	116	0,30	0,59	2,96	0,09	17,79	$5{\cdot}10^{-5}$	7,05	**0,009**[b,d]
SK04-M	Suche	B04	116	63,66	**$1{\cdot}10^{-12}$**[c]	18,79	**$3{\cdot}10^{-5}$**[b,c]	18,09	$4{\cdot}10^{-5}$	0,19	0,67
		B05	72	29,31	**$9{\cdot}10^{-7}$**[c]	15,56	**$2{\cdot}10^{-4}$**[b,c]	12,44	$8{\cdot}10^{-4}$	1,86	0,18
		B06	96	21,10	**$1{\cdot}10^{-5}$**[c]	5,93	**0,02**[b,c]	11,63	0,001	0,05	0,83
		B12	56	11,02	**0,002**[c]	10,43	**0,002**[b,c]	4,96	0,03	0,73	0,40
		B16	116	71,31	**$1{\cdot}10^{-13}$**[c]	18,15	**$4{\cdot}10^{-5}$**[b,c]	17,12	$7{\cdot}10^{-5}$	0,83	0,36
		B17	60	27,47	**$3{\cdot}10^{-6}$**[c]	10,73	**0,002**[b,c]	9,99	0,003 [d]	4,57	0,04
		B18	84	25,02	**$3{\cdot}10^{-6}$**[c]	5,85	**0,02**[b,d]	9,85	0,002 [c]	0,48	0,49
		M45	116	11,21	**0,001**[c]	4,77	**0,03**[b,c]	10,07	0,002	0,38	0,54
SK07-M[e]	Benutzerfreundlichkeit	B01	40	$2{\cdot}10^{-6}$	1,00	0,97	0,33	2,26	**0,14**[a]	0,48	0,49
		B04	116	21,76	**$9{\cdot}10^{-6}$**[c]	13,36	**$4{\cdot}10^{-4}$**[b,c]	12,17	$7{\cdot}10^{-4}$ [c]	0,94	0,34
		B06	96	20,29	**$2{\cdot}10^{-5}$**[c]	9,35	**0,003**[b,d]	9,48	0,003	1,52	0,22
		B12	56	4,12	**0,05**[c]	16,42	**$2{\cdot}10^{-4}$**[b,c]	6,30	0,02	2,09	0,15
		B16	116	31,77	**$1{\cdot}10^{-7}$**[c]	11,42	**0,001**[b,c]	13,99	$3{\cdot}10^{-4}$ [c]	1,70	0,19
		B17	60	11,14	**0,002**[c]	14,69	**$3{\cdot}10^{-4}$**[b,c]	11,92	0,001 [c]	3,73	0,06
		B18	84	15,66	**$2{\cdot}10^{-4}$**[c]	7,73	**0,007**[b,c]	10,30	0,002 [c]	0,77	0,38

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.

[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

[e] Entspricht auch der Skala SK18-M.

[f] Entspricht auch den Skalen SK15-M und SK19-M.

[g] Entspricht auch der Skala SK13-M.

Tab. E.66 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
SK08-M	Suche	B12	56	15,65	**$2{\cdot}10^{-4}$**[c]	8,00	**0,007**[b,c]	4,86	0,03	0,63	0,43
		B16	116	87,09	**$1{\cdot}10^{-15}$**[c]	17,89	**$5{\cdot}10^{-5}$**[b,c]	23,76	$4{\cdot}10^{-6}$ [c]	1,69	0,20
		B18	84	32,06	**$2{\cdot}10^{-7}$**[c]	7,16	**0,009**[b,d]	12,85	$6{\cdot}10^{-4}$	1,20	0,28
SK09-M	Benutzerfreundlichkeit	B04	116	28,81	**$4{\cdot}10^{-7}$**[c]	12,25	**$7{\cdot}10^{-4}$**[b,c]	18,72	$3{\cdot}10^{-5}$	2,61	0,11
		B06	96	14,91	**$2{\cdot}10^{-4}$**[c]	6,00	**0,02**[b,d]	18,32	$5{\cdot}10^{-5}$	3,60	0,06
		B12	56	3,15	0,08	13,39	**$6{\cdot}10^{-4}$**[b,c]	16,28	$2{\cdot}10^{-4}$	2,79	0,10
		B16	116	31,66	**$1{\cdot}10^{-7}$**[c]	12,17	**$7{\cdot}10^{-4}$**[b,c]	27,27	$8{\cdot}10^{-7}$ [c]	4,58	0,03
		B17	60	10,43	**0,002**[c]	9,57	**0,003**[b,d]	13,09	$6{\cdot}10^{-4}$ [c]	5,60	0,02
		B17	60	8,29	**0,006**[c]	6,27	0,02	9,59	0,003	8,85	**0,004**[b,c]
		B18	84	18,12	**$6{\cdot}10^{-5}$**[c]	8,81	**0,004**[b,c]	19,05	$4{\cdot}10^{-5}$	4,83	0,03
SK11-M[f]	Eigenleistung	B06	96	16,25	**$1{\cdot}10^{-4}$**[c]	0,36	0,55	2,73	**0,10**[a]	0,005	0,94
		B10	76	18,60	**$5{\cdot}10^{-5}$**[c]	0,94	0,33	1,22	**0,27**[a]	0,001	0,97
		B17	60	28,89	**$2{\cdot}10^{-6}$**[c]	0,64	0,43	2,82	**0,10**[a]	0,19	0,66
		B18	84	18,03	**$6{\cdot}10^{-5}$**[c]	0,01	0,92	2,33	**0,13**[a]	0,57	0,45
SK12-F	Suchergebnis	B11	24	5,16	**0,03**[d]	1,10	0,31	1,75	**0,20**[a]	0,12	0,74
		B16	116	81,84	**$6{\cdot}10^{-15}$**[c]	16,20	**$1{\cdot}10^{-4}$**[b,c]	26,37	$1{\cdot}10^{-6}$ [c]	3,20	0,08
SK12-M	Suchergebnis	B12	56	13,60	**$5{\cdot}10^{-4}$**[c]	7,30	**0,009**[b,d]	8,44	0,005	0,18	0,68
SK13-F	Benutzerfreundlichkeit	B01	40	0,09	0,77	1,31	0,26	1,22	**0,28**[a]	2,51	0,12
		B17	60	0,16	0,69	0,14	0,71	6,46	0,01	8,80	**0,004**[b,c]
SK-E[g]	Ease of Use (EUCS)	B01	40	0,72	0,40	0,002	0,96	2,07	**0,16**[a]	0,28	0,60
		B17	60	0,94	0,34	0,23	0,64	7,01	0,01	6,51	**0,01**[b,d]
SK14-F	Suche	B04	116	35,15	**$4{\cdot}10^{-8}$**[c]	16,94	**$7{\cdot}10^{-5}$**[b,c]	13,81	$3{\cdot}10^{-4}$	1,11	0,29
		B05	72	9,60	**0,003**[c]	5,39	**0,02**[b,d]	2,84	0,10	2,23	0,14
		B05	72	13,27	**$5{\cdot}10^{-4}$**[c]	1,57	0,21	1,91	**0,17**[a]	0,005	0,94
		B06	96	22,54	**$8{\cdot}10^{-6}$**[c]	8,03	**0,006**[b,c]	11,66	0,001	1,55	0,22
		B11	24	9,94	**0,005**[d]	7,17	0,01	3,11	**0,09**[a]	4,47	0,05
		B12	56	7,45	**0,009**[c]	10,65	**0,002**[b,c]	5,76	0,02 [d]	1,87	0,18
		B16	116	50,72	**$1{\cdot}10^{-10}$**[c]	15,31	**$2{\cdot}10^{-4}$**[b,c]	13,15	$4{\cdot}10^{-4}$	0,43	0,51
		B18	84	24,61	**$4{\cdot}10^{-6}$**[c]	9,35	**0,003**[b,c]	10,18	0,002	1,16	0,28
		M45	116	5,55	**0,02**[c]	7,18	**0,009**[b,c]	10,78	0,001	0,90	0,34
		S05	104	0,19	0,66	8,85	**0,004**[b,c]	13,70	$4{\cdot}10^{-4}$	1,81	0,18

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.

[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

[e] Entspricht auch der Skala SK18-M.

[f] Entspricht auch den Skalen SK15-M und SK19-M.

[g] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.66 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
SK14-M	Suche	B04	116	37,02	**$2{\cdot}10^{-8}$**[c]	15,95	**$1{\cdot}10^{-4}$**[b,c]	18,43	$4{\cdot}10^{-5}$ [c]	1,39	0,24
		B05	72	13,20	**$5{\cdot}10^{-4}$**[c]	6,46	**0,01**[b,d]	3,21	0,08	1,92	0,17
		B06	96	25,90	**$2{\cdot}10^{-6}$**[c]	11,33	**0,001**[b,c]	12,30	$7{\cdot}10^{-4}$ [c]	1,98	0,16
		B10	76	14,34	**$3{\cdot}10^{-4}$**[c]	7,58	**0,007**[b,c]	9,03	0,004	0,01	0,91
		B12	56	8,00	**0,007**[c]	11,14	**0,002**[b,c]	6,47	0,01	0,95	0,33
		B16	116	58,44	**$8{\cdot}10^{-12}$**[c]	14,19	**$3{\cdot}10^{-4}$**[b,c]	16,35	$1{\cdot}10^{-4}$ [c]	0,55	0,46
		B17	60	17,76	**$9{\cdot}10^{-5}$**[c]	10,16	**0,002**[b,c]	8,29	0,006	2,46	0,12
		B17	60	24,29	**$8{\cdot}10^{-6}$**[c]	5,32	0,02	3,20	**0,08**[a]	1,77	0,19
		B18	84	29,93	**$5{\cdot}10^{-7}$**[c]	10,15	**0,002**[b,c]	11,27	0,001 [c]	1,15	0,29
SK15-F	Eigenleistung	B03	28	0,10	0,75	0,49	0,49	9,94	**0,004**[b,c]	0,005	0,94
		B04	116	31,46	**$2{\cdot}10^{-7}$**[c]	0,49	0,48	6,77	**0,01**[b,c]	0,04	0,85
		B05	72	37,65	**$5{\cdot}10^{-8}$**[c]	0,55	0,46	8,30	**0,005**[b,d]	0,27	0,60
		M10	116	13,90	**$3{\cdot}10^{-4}$**[c]	0,23	0,63	8,63	**0,004**[b,c]	0,04	0,83
		M16	116	10,14	**0,002**[c]	0,13	0,72	9,16	**0,003**[b,c]	0,02	0,90
		M37	116	12,67	**$5{\cdot}10^{-4}$**[c]	$1{\cdot}10^{-4}$	0,99	10,43	**0,002**[b,c]	0,50	0,48
		M45	116	7,82	**0,006**[c]	0,66	0,42	8,55	**0,004**[b,c]	0,04	0,84
		S01	116	0,07	0,79	0,02	0,90	10,29	**0,002**[b,c]	0,02	0,88
		S02	116	0,32	0,57	0,002	0,97	11,78	**$8{\cdot}10^{-4}$**[b,c]	0,007	0,93
		S03	116	0,67	0,41	$3{\cdot}10^{-4}$	0,99	7,27	**0,008**[b,c]	0,08	0,77
		S04	116	3,69	0,06	0,003	0,96	11,24	**0,001**[b,c]	0,35	0,55
		S05	104	7,67	**0,007**[c]	0,25	0,62	6,82	**0,01**[b,d]	0,08	0,78
		S06	116	15,58	**$1{\cdot}10^{-4}$**[c]	0,29	0,59	10,32	**0,002**[b,c]	1,23	0,27
SK16-F	Aufgabe	B01	40	2,77	0,10	0,79	0,38	1,13	**0,30**[a]	1,16	0,29
		B03	28	2,79	0,11	3,74	0,07	2,42	**0,13**[a]	0,22	0,65
		B05	72	14,30	**$3{\cdot}10^{-4}$**[c]	3,73	0,06	3,48	**0,07**[a]	0,79	0,38
		B06	96	14,93	**$2{\cdot}10^{-4}$**[c]	1,00	0,32	2,33	**0,13**[a]	0,75	0,39
		B11	24	7,74	**0,01**[c]	1,22	0,28	0,15	**0,71**[a]	0,80	0,38
		B12	56	10,80	**0,002**[c]	6,87	**0,01**[b,d]	1,49	**0,23**[a]	0,44	0,51
		B18	84	14,61	**$3{\cdot}10^{-4}$**[c]	0,68	0,41	3,21	**0,08**[a]	1,25	0,27
SK16-M	Aufgabe	B12	56	8,51	**0,005**[c]	7,47	**0,009**[b,d]	4,40	0,04	0,71	0,40
		B16	116	51,13	**$1{\cdot}10^{-10}$**[c]	6,19	**0,01**[b,c]	15,33	$2{\cdot}10^{-4}$ [c]	0,29	0,59
SK17-F	Suche	B11	24	7,78	**0,01**[c]	0,74	0,40	1,72	**0,21**[a]	0,008	0,93
		B16	116	76,24	**$3{\cdot}10^{-14}$**[c]	16,01	**$1{\cdot}10^{-4}$**[b,c]	22,58	$6{\cdot}10^{-6}$	2,21	0,14
SK17-M	Suche	B06	96	34,73	**$6{\cdot}10^{-8}$**[c]	7,19	**0,009**[b,c]	8,56	0,004	1,59	0,21
		B18	84	30,09	**$5{\cdot}10^{-7}$**[c]	6,94	**0,01**[b,c]	12,36	$7{\cdot}10^{-4}$	1,45	0,23

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.

[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

[e] Entspricht auch der Skala SK18-M.

[f] Entspricht auch den Skalen SK15-M und SK19-M.

[g] Entspricht auch der Skala SK13-M.

Fortsetzung auf nächster Seite

Tab. E.66 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
SK18-F	Benutzerfreundlichkeit	B03	28	0,07	0,80	5,92	0,02	2,00	**0,17**[a]	2,39	0,14
		B04	116	13,68	**$3{\cdot}10^{-4}$**[c]	9,44	**0,003**[b,c]	10,55	0,002	0,68	0,41
		B05	72	0,74	0,39	9,64	**0,003**[b,c]	3,38	0,07	2,67	0,11
		B05	72	3,09	0,08	5,17	0,03	1,22	**0,27**[a]	0,12	0,73
		B06	96	14,44	**$3{\cdot}10^{-4}$**[c]	9,23	**0,003**[b,d]	8,70	0,004	1,13	0,29
		B10	76	4,65	**0,03**[d]	4,71	**0,03**[b,d]	7,90	0,006 [c]	1,45	0,23
		B11	24	1,62	0,22	5,20	0,03	0,31	**0,58**[a]	0,12	0,74
		B12	56	2,09	0,15	11,94	**0,001**[b,c]	5,56	0,02	0,76	0,39
		B16	116	19,22	**$3{\cdot}10^{-5}$**[c]	10,09	**0,002**[b,c]	12,17	$7{\cdot}10^{-4}$ [c]	1,05	0,31
		B17	60	3,35	0,07	2,55	0,12	0,95	**0,33**[a]	1,96	0,17
		B18	84	8,40	**0,005**[c]	8,67	**0,004**[b,c]	9,24	0,003 [c]	0,99	0,32
SK19-F	Eigenleistung	B05	72	22,62	**$1{\cdot}10^{-5}$**[c]	0,41	0,52	3,49	**0,07**[a]	0,004	0,95
		B06	96	24,01	**$4{\cdot}10^{-6}$**[c]	0,30	0,59	1,76	**0,19**[a]	0,18	0,67
		B10	76	13,83	**$4{\cdot}10^{-4}$**[c]	1,05	0,31	1,13	**0,29**[a]	$9{\cdot}10^{-4}$	0,98
		B16	116	22,58	**$6{\cdot}10^{-6}$**[c]	0,25	0,62	3,42	**0,07**[a]	1,10	0,30
		B17	60	21,19	**$2{\cdot}10^{-5}$**[c]	0,18	0,67	2,37	**0,13**[a]	0,36	0,55
		B18	84	7,76	**0,007**[c]	0,15	0,70	2,70	**0,10**[a]	0,09	0,76
E01	Ich glaube, ich werde in zehn Minuten ... relevante Dokumente finden.	B18	84	16,88	**$1{\cdot}10^{-4}$**[c]	0,69	0,41	5,90	**0,02**[b,d]	0,73	0,40
		S01	116	0,86	0,36	2,30	0,13	7,42	**0,007**[b,c]	1,11	0,29
		S05	104	13,32	**$4{\cdot}10^{-4}$**[c]	0,29	0,59	6,68	**0,01**[b,c]	0,04	0,84
E02	Wie wahrscheinlich ist es, dass diese Suchmaschine Ihnen dabei helfen wird eine gute Leistung zu erbringen?	B06	96	23,23	**$6{\cdot}10^{-6}$**[c]	7,66	**0,007**[b,c]	8,45	0,005	0,02	0,89
		B11	24	5,55	**0,03**[c]	11,20	0,003	0,22	**0,64**[a]	0,96	0,34
		B12	56	14,07	**$5{\cdot}10^{-4}$**[c]	11,11	**0,002**[b,c]	3,04	0,09	0,24	0,62
		B12	56	10,71	**0,002**[c]	8,14	0,006	1,91	**0,17**[a]	0,04	0,84
		B16	116	39,41	**$7{\cdot}10^{-9}$**[c]	10,10	**0,002**[b,c]	17,35	$6{\cdot}10^{-5}$	0,009	0,93
		B18	84	24,35	**$4{\cdot}10^{-6}$**[c]	7,52	**0,008**[b,d]	10,21	0,002	0,03	0,87
E03	Wie wahrscheinlich ist es, dass Sie mithilfe dieser Suchmaschine zu einem schnellen Ergebnis kommen?	B11	24	$7{\cdot}10^{-7}$	1,00	0,68	0,42	0,68	**0,42**[a]	0,88	0,36
E04	Wie wahrscheinlich ist es, dass Sie von der Leistung, die Sie mithilfe dieser Suchmaschine erbringen, sehr überzeugt sind?	B01	40	0,63	0,43	0,72	0,40	1,27	**0,27**[a]	0,94	0,34
		B03	28	0,02	0,89	4,77	0,04	1,09	**0,31**[a]	0,64	0,43
		B04	116	30,68	**$2{\cdot}10^{-7}$**[c]	9,04	**0,003**[b,c]	10,36	0,002 [c]	1,80	0,18
		B05	68	7,52	**0,008**[c]	10,67	**0,002**[b,c]	6,65	0,01 [d]	0,74	0,39
		B11	24	0,99	0,33	11,56	**0,003**[b,d]	2,06	0,17	1,22	0,28
		B11	24	3,65	0,07	4,99	0,04	0,24	**0,63**[a]	0,33	0,57
		B16	116	42,48	**$2{\cdot}10^{-9}$**[c]	9,59	**0,002**[b,c]	9,65	0,002	0,75	0,39
		B17	56	18,74	**$7{\cdot}10^{-5}$**[c]	3,71	0,06	0,44	**0,51**[a]	0,91	0,34

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.

Fortsetzung auf nächster Seite

[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.

[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.

[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.

[e] Entspricht auch der Skala SK18-M.

[f] Entspricht auch den Skalen SK15-M und SK19-M.

[g] Entspricht auch der Skala SK13-M.

Tab. E.66 (Fortsetzung)

				Kovariate		System		Erwartung		Interaktion	
ID	**Beschreibung**	**Kov.**	**n**	**F**	**p**	**F**	**p**	**F**	**p**	**F**	**p**
E05	Wie wahrscheinlich ist es, dass Sie mit dieser Suchmaschine sehr zufrieden sind?	B01	40	$1{\cdot}10^{-7}$	1,00	1,00	**0,32**[a]	4,01	0,05	$2{\cdot}10^{-8}$	1,00
		B01	40	0,004	0,95	7,35	0,01	3,68	**0,06**[a]	0,61	0,44
		B11	24	4,69	**0,04**[d]	2,82	**0,11**[a]	2,10	0,16	1,12	0,30
		B11	24	10,12	**0,005**[d]	3,83	0,07	0,55	**0,47**[a]	0,01	0,92
		M16	116	8,17	**0,005**[c]	2,88	**0,09**[a]	13,22	$4{\cdot}10^{-4}$	1,29	0,26
E06-M	Erwartungsskala	B04	116	33,00	$\mathbf{8{\cdot}10^{-8}}$[c]	13,13	$\mathbf{4{\cdot}10^{-4}}$[b,c]	19,22	$3{\cdot}10^{-5}$	0,41	0,52
		B06	96	12,11	$\mathbf{8{\cdot}10^{-4}}$[c]	8,41	**0,005**[b,c]	11,37	0,001	2,08	0,15
		B10	76	17,87	$\mathbf{7{\cdot}10^{-5}}$[c]	10,78	**0,002**[b,d]	15,87	$2{\cdot}10^{-4}$ [c]	0,02	0,88
		B11	24	0,03	0,87	0,72	0,41	0,46	**0,51**[a]	0,06	0,82
		B16	116	43,74	$\mathbf{1{\cdot}10^{-9}}$[c]	12,24	$\mathbf{7{\cdot}10^{-4}}$[b,c]	22,54	$6{\cdot}10^{-6}$	0,61	0,44
		B18	84	15,06	$\mathbf{2{\cdot}10^{-4}}$[c]	8,83	**0,004**[b,c]	18,69	$4{\cdot}10^{-5}$ [c]	1,63	0,21
		M45	116	6,54	**0,01**[d]	4,72	**0,03**[b,d]	15,36	$2{\cdot}10^{-4}$	0,51	0,48

[a] Dieser Effekt entfällt im Vergleich zur Varianzanalyse.
[b] Dieser Effekt kommt im Vergleich zur Varianzanalyse neu hinzu.
[c] Dieser Effekt ist im Rahmen der Kovarianzanalyse eindeutig signifikant.
[d] Dieser Effekt ist im Rahmen der Kovarianzanalyse in der Tendenz signifikant.
[e] Entspricht auch der Skala SK18-M.
[f] Entspricht auch den Skalen SK15-M und SK19-M.
[g] Entspricht auch der Skala SK13-M.